DOCUMENT RESUME

ED 303 783 CS 009 523

AUTHOR Haertel, Edward; And Others
TITLE Report of the NAEP Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons. Technical Report.
INSTITUTION National Center for Education Statistics (ED), Washington, DC.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CS-89-499
PUB DATE Jan 89
NOTE 250p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC10 Plus Postage.
DESCRIPTORS *Educational Assessment; *Educational Trends; Elementary Secondary Education; *Mathematics Achievement; National Competency Tests; National Norms; *Reading Achievement
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

This technical report, the final product of the deliberations of a panel charged with an external review of the National Assessment of Educational Progress (NAEP), considers three broad issues: (1) the apparent lack of comparability between the findings of the 1984 and 1986 reading assessments (reading anomaly), (2) the accuracy of NAEP trend data, particularly in reading and mathematics, and apparent inconsistencies between NAEP trend data and those from other major tests; and (3) problems and possible solutions in the expansion of NAEP to include a state-by-state assessment. Findings indicated that the bulk of the apparent declines in 9- and 17-year-olds' reading scores was probably artifactual; that while NAEP is a better barometer of national achievement trends than any available alternative, the quality of its trend reporting could be improved considerably (three recommendations are given); and that state-level assessments should be managed by a separate program unit within the National-NAEP organization, and should be parallel to the National-NAEP in most respects. The report concludes with 13 papers (individually authored or coauthored by panel members) addressing particular issues within the charge of the panel. (SR)

NATIONAL CENTER FOR EDUCATION STATISTICS

Technical Report January 1989

Report of the NAEP Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons

U.S. Department of Education
Office of Educational Research and Improvement CS 89-499

NATIONAL CENTER FOR EDUCATION STATISTICS

Technical Report January 1989

Report of the NAEP Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons

Edward Haertel, Chair

Pascal D. Forgione, Jr., Chair
Subpanel on State Comparisons

Herbert J. Walberg, Chair
Subpanel on Anomaly and Trends

Janet Baldwin
R. Darrell Bock
Leigh Burstein
Dale Carlson
Jeanne S. Chall
John T. Guthrie
Larry V. Hedges
Dan Melnick
Mark D. Musick
Tej Pandey
William H. Schmidt
David E. Wiley

U.S. Department of Education
Office of Educational Research and Improvement CS 89-499

National Center for Education Statistics

"The purpose of the Center shall be to collect, and analyze, and disseminate statistics and other data related to education in the United States and in other nations."—Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e–1).

Foreword

History and Charge of Panel

In December, 1987, a Technical Review Panel was formed by the Center for Education Statistics to conduct an external review of the National Assessment of Educational Progress (NAEP). The panel was charged with examining three broad issues:

--The apparent lack of comparability between the findings of the 1984 and 1986 reading assessments;

--The accuracy of NAEP trend data, particularly in reading and mathematics, and apparent inconsistencies between NAEP trend data and those from other major tests; and

--Problems and possible solutions in the expansion of NAEP to include a State-by-State assessment.

The panel was organized into two subpanels to carry out its investigations. One subpanel, chaired by Herbert J. Walberg, addressed the issues of the 1986 reading anomaly and of the accuracy of NAEP trends. The other, chaired by Pascal D. Forgione, Jr., addressed issues in the expansion of NAEP to permit State-level reporting and comparisons. In addition to separate deliberations by the two subpanels, the entire group met to exchange views on all three issues, and to reach agreement on its recommendations and conclusions. The chair for the entire panel was Edward Haertel.

The panel held 2-day meetings in December, January, and February, and a final, 1-day meeting late in April. Based on discussions, data provided by the Educational Testing Service (ETS), interviews with ETS personnel, and other information, the panel formulated a set of recommendations for the conduct of the NAEP, designed to minimize the probability of a recurrence of the reading anomaly, to assure the accuracy and continuity of NAEP trends, and to address concerns that arise in the expansion of NAEP to provide State-level achievement estimates and comparisons.

Most members of the panel contributed individually authored papers addressing particular issues within the charge of the panel. Two of the panel members, Dr. Forgione and Dr. Guthrie, coauthored their papers with colleagues who were not members of the panel. To varying degrees, these papers reflect the results of discussions and deliberations by the panel as a whole, but each represents primarily a single author's position. This is as it should be. The members of the panel were deliberately chosen for their varied areas of expertise and their varied perspectives. and in their respective papers, each addressed areas in which she or he was especially well qualified.

Following the preparation of the separate papers, Dr. Haertel reviewed their findings and recommendations, and drafted the "Report of the Panel," which summarizes and supports the findings presented in the separately authored papers. This draft was circulated to all of the panel members, and revised in response to the comments received. The Report addresses each of the panel's three charges in turn, summarizing major points from all of the relevant background papers, and from the panel's deliberations. At many points, the reader is referred back to the separate papers for more extended discussion.

This document is the final product of the panel's deliberations. It includes an executive summary, the panel's recommendations and conclusions, an appendix in which one panel member qualifies her endorsement of these recommendations and conclusions, the "Report of the Panel," and the separately authored papers.

Table of Contents

Letter of Transmittal

June 27, 1988

Mr. Emerson Elliott
Acting Commissioner
National Center for Education Statistics
Room 400
555 New Jersey Avenue NW
Washington, DC 20208-1405

Dear Mr. Elliott:

I am pleased to transmit to you the report of the Technical Review Panel on the National Assessment of Educational Progress.

As NAEP evolves to provide new kinds of information, it is critical that its findings be trustworthy and reliable, that the continuity of its trend lines be preserved, and that the State-level information it will provide be fair, credible, and maximally useful.

Our panel has worked to develop a coherent vision of a National Assessment that would meet these challenges, and we hope that our report is of value to you as you plan for the future.

We appreciate the support and interest in our work shown by Assistant Secretary Chester E. Finn, Jr., by you, and by your excellent staff at the National Center for Education Statistics. We also acknowledge the help and cooperation of the NAEP staff at the Educational Testing Service.

Sincerely,

Edward Haertel
Chair

Members of the NAEP Technical Review Panel

Edward Haertel
Stanford University
Chairman

Subpanel on Anomaly and Trends

Herbert J. Walberg, Chair
University of Illinois at Chicago

Janet Baldwin
American Council on Education

Jeanne S Chall
Harvard University

Larry V. Hedges
University of Chicago

Tej Pandey
California State Department of Education

William H. Schmidt
National Science Foundation

David E. Wiley
Northwestern University

Subpanel on State Comparisons

Pascal D. Forgione, Jr., Chair
Connecticut State Department of Education

R. Darrell Bock
University of Chicago and NORC

Leigh Burstein
University of California, Los Angeles

Dale Carlson
California State Department of Education

John T. Guthrie
University of Maryland

Dan Melnick
Joint Economic Committee

Mark D. Musick
Southern Regional Education Board

Executive Summary

The NAEP Technical Review Panel was convened in December 1987, by the Center for Education Statistics and charged with examining three broad issues:

--The apparent lack of comparability between the findings of the 1984 and 1986 reading assessments (reading anomaly);

--The accuracy of NAEP trend data, particularly in reading and mathematics, and apparent inconsistencies between NAEP trend data and those from other major tests; and

--Problems and possible solutions in the expansion of NAEP to include a State-by-State assessment.

The panel has produced a set of joint recommendations and conclusions and a report summarizing its deliberations. These are supported by individual papers on particular topics. The panel reached consensus on its recommendations, conclusions, and report, with the exception that one panel member had minor reservations concerning one recommendation and one conclusion (see "Qualification of Endorsement and Recommendations" on pp. 11-12).

The 1986 Reading Anomaly

While acknowledging that real declines in reading ability may have occurred, the panel was nearly unanimous in concluding that the bulk of the apparent declines in 9- and 17-year-old reading scores was probably artifactual. In reaching this conclusion, the panel concurred with the Technical Report on the anomaly by Educational Testing Service. The panel generally endorsed the ETS investigation of the anomaly, but criticized the almost exclusive focus on declines in mean scores. More attention should have been paid to the substantial increases from 1984 to 1986 in the variances of score distributions at all three age/grade levels. The panel also criticized the ETS report for considering possible hypotheses in isolation from one another, when a combination of two or more might easily have explained the bulk of the observed score declines. The panel also suggested some additional hypotheses that may merit consideration.

The Accuracy of NAEP Trends

The panel concluded that despite its imperfections, NAEP is a better barometer of national achievement trends than any available alternative. The only other national, longitudinal achievement data collected over a comparable span of years come from college admissions tests, and these cover relatively limited domains of content, test only at the high school level, and examine self-selected groups of students that are not nationally representative of their age cohorts. At the same time, the panel concluded that the quality of NAEP trend reporting could be improved considerably.

The panel's three principal recommendations for improving the accuracy and authoritativeness of NAEP trends were first, to assure greater consistency over assessment cycles in the objectives covered; second, to assure greater care in revising the format of assessment materials or testing sessions, or any other aspects of NAEP procedures that might impact the continuity of NAEP trends; and third, to provide for an ongoing statistical evaluation and audit of NAEP data collection and reported findings, independent of the NAEP contractor.

State-Level NAEP

State-level assessments should be managed by a separate program unit within the National-NAEP organization, and should be parallel to the National-NAEP in most respects. In designing State-NAEP procedures, comparability among States and between State-level and national data is paramount. This implies a centralized administration plan. State samples should be drawn by the same contractor responsible for National-NAEP samples, and the National-NAEP organization should be responsible for training State-NAEP examiners, probably personnel provided by States for the 6- to 8-week period of training and data collection. The 1990 and 1992 pilots authorized by the Hawkins-Stafford law should be used to explore alternative administration procedures. The panel recommended an expanded NAEP data collection, covering more learning outcomes and more background questions. Results at both national and State levels should be reported at a greater level of specificity. Sufficient data should be collected from each student to permit accurate estimation of individual scores (although anonymity would be preserved). The amount of individual student time devoted to NAEP should be expanded to accommodate these changes. State-level results should be released promptly. In addition to reports on absolute levels of achievement, State results should be referenced to the performance of comparable States, national samples of students matched to State characteristics, or in other ways that account for demographic differences. A variety of comparison methods should be explored and reported in 1990 and 1992.

Part 1. Recommendations and Conclusions

NAEP Technical Review Panel
Recommendations and Conclusions

May 21, 1988

Recommendation 1

The frameworks that have been used to organize NAEP objectives are inadequate in terms of comprehensiveness, specificity, and stability over assessment cycles. Knowledge and skills assessed should be drawn from an explicit, comprehensive, detailed, and stable domain. Content changes over assessment cycles should be specified in terms of this domain, and should be undertaken only after careful evaluation.

Each organizing domain would include descriptions of the knowledge, skills, or other possible learning outcomes that might be intended in a content area such as reading or mathematics. Domains would almost certainly encompass more than the range of learning outcomes represented by present NAEP exercise pools, and there would be no expectation that future assessments would necessarily attempt coverage of all of the particular learning outcomes within a domain. Domains would provide the basis for the more detailed reporting of NAEP results proposed under Recommendation 8.

NAEP exercises would be referenced to these domains. However, the domains would not be simple classification schemes for exercises, nor would they specify particular forms of test items corresponding to different learning outcomes. Test item responses are a consequence of more than one skill or ability, some intended and some not. The linkage of NAEP exercises to a skill domain will involve significant issues of item validity, including exercise formats appropriate for respondents at different ages, and more generally, the ancillary skills exercises may require. These ancillary skill requirements can reduce the validity of exercises as indicators of the learning outcomes they are designed to measure.

Recommendation 2

NAEP should broadly assess important learning outcomes. This implies increased emphasis on higher-level learning outcomes now considered critical for all students. Priorities should be guided by expert subject-matter perspectives, as well as current substantive and methodological research.

A major purpose of NAEP is to inform and focus discussion of education policy and practice. It follows that, at both the national and State levels, NAEP must represent a full and rich conception of important cognitive learning outcomes in the domains assessed. These include the learnings covered by the typical curricula of the Nation's schools but should reach beyond the typical. The content of NAEP must not be reduced to the intersection of the several States' curriculum frameworks, nor to any "lowest common denominator." Assessment objectives should reflect the best

thinking of subject-matter specialists, and should focus on emerging views of learning and knowledge in all content subjects, such as history and literature. The NAEP objectives should lay a solid foundation for discussion among professionals and practitioners about such issues as curriculum. teaching, inservice education, school organization, and policy alternatives available to State leaders.

The design of NAEP exercises should capitalize on methodological advances to assure the valid assessment of complex, higher level learning outcomes as well as important factual knowledge, basic skills, and other lower level outcomes usually achieved earlier and considered prerequisite for higher level learnings. Attention should be given to the measurement of processes such as reading, writing, and problem solving in the context of content subjects. Finally, various modes of assessment should provide ample opportunity for students to display their productive abilities. This will require increased emphasis on writing, speaking, and interacting in both real-world and school tasks.

Recommendation 3

State-level NAEP should collect information on student and teacher background factors and on schooling processes, as well as achievement. In particular, a core set of student background questions should be included on both national-level NAEP and State-level NAEP. State-level NAEP should also explore the feasibility of collecting information on opportunity to learn and other schooling processes possibly linked to achievement. Questions should be used over a series of assessments so that trends can be observed.

NAEP achievement data become more meaningful and more useful if they can be linked to carefully chosen school, teacher, and student characteristics and schooling processes, as well as community, home, and family characteristics, and students' out-of-school activities. Judicious selection of background questions is essential. Priority should be given to those that are (a) important for describing patterns in NAEP achievement data, (b) plausibly related to achievement, or (c) reflective of other valued schooling outcomes.

Background questions may be addressed to students, teachers, and principals. In general, the same questions should be used in State-level NAEP and in national-level NAEP, and arbitrary changes in background questions from one assessment cycle to the next should be avoided. Changes may be necessary to assure coverage of schooling process variables important to particular content areas assessed concurrently. Information about changes in curriculum and instructional methods can be critical to the interpretation of NAEP trends. If a reliable, efficient. and unobtrusive method can be found for collecting data on students' opportunity to learn the content of concurrently administered achievement exercises, this information might be especially valuable.

Recommendation 4

The amount of individual student time devoted to the conduct of NAEP should be expanded.

Recommendations 2 and 3 call for increased amounts of information about individual examinees. In order to accomplish this, increased time is required for responding to achievement exercises and survey questions on student background and experiences.

Recommendation 5

Assessment procedures impacting the continuity of national trends should not be altered unless there is a compelling reason to do so. Changes should be made only after systematic examination of the likely consequences and justification in terms of NAEP priorities. Old and new procedures should be carried in parallel for at least one assessment cycle, on a scale sufficient to assure continuity of national trends.

In The Nation's Report Card, Alexander and James place the highest priority on the maintenance of continuity in the trend lines of NAEP achievement assessments (p. 7). We strongly concur with this priority. Undoubtedly, there will be profound changes in future NAEP data collections, especially in light of recommendations to extend NAEP to permit State-by-State comparisons. However, whatever modifications are made in the overall program design, it is mandatory that the procedures used to collect the data for national trend estimates be parallel in every important respect. During transitions when old and new procedures are carried in parallel, not only the assessment exercises themselves but also the data collection procedures should remain the same. Only this will assure that scale scores with the same meaning as those available before 1986 can be calculated for 1988 and beyond.

We strongly endorse the current (1988) ETS replications of the 1984 and 1986 procedures as a vital source of information for future design decisions. However, we do not believe that ad hoc investigations are sufficient to assure the continuity of NAEP time series. A new process should be developed to assure adequate and systematic evaluation of proposed procedural changes. The technical advisory process to NAEP should comprehensively incorporate considerations of procedural design and audit, as well as sample design and analysis. This implies formal review of on-site administration conditions and procedures, instructions and student conformity to them, etc., as well as timing and booklet design. This also implies that the technical advisory body(ies) should be composed of individuals representing all relevant areas of expertise.

Recommendation 6

NAEP data must be collected so as to assure comparability across States. Sampling procedures, core instrumentation, and conditions of administration must be uniform. Although States may choose to augment their data collection, the minimum design must be sufficient to provide comparable

estimates of achievement levels and distributions for each State. A centralized administration plan will best serve the ends of comparability. As part of the pilot assessments authorized by the Hawkins-Stafford law, NCES should study the effects of alternative administration procedures on comparability.

State-level NAEP samples should be drawn following the same procedures as for the national sample, with the possible exception that schools may be drawn as the primary sampling units rather than being clustered within counties or groups of counties. The present practice of returning questionnaires on excluded students should be continued, and the percent of students excluded should be reported along with State assessment results.

For those States that wish to participate in the assessment more extensively, one or more options should be developed for an expanded assessment, which might include an expanded student sample, additional background or achievement questions, or both. The national core instrumentation must precede any supplementation.

Recommendation 7

Scores of individual students should be estimated and made available for analysis. However, consistent with confidentiality restrictions in the law, particular students shall not be identified.

Accurate score estimates for individual examinees would permit NAEP to report the estimated score distributions referred to in Recommendation 9 below, and would greatly simplify other analyses by the NAEP contractor, as well as secondary users. In recent National Assessment data collections, the way in which the Balanced Incomplete Block (BIB) spiraled booklet assignments were designed made it impossible to generate accurate score estimates for individual examinees using conventional Item Response Theoretic (IRT) procedures. In future assessments, the data obtained from students should permit the accurate estimation of their individual performance levels. Individual scores could be obtained using either a BIB-Spiraled design or alternative designs.

Recommendation 8

NAEP results should be reported at a greater level of specificity. This reporting should permit distinctions among important parts of the domain. This will imply the use of multiple scores or scales within domains.

The skill domains discussed in Recommendation 1 must comprise important and conceptually distinct core components. NAEP should, in terms of these components, identify the subdomains on which trends in achievement will be reported. In the past, although a consensus approach for defining objectives has been followed for each assessment, little attention appears to have been paid to the continuity of consensus over time. Appropriate subdomain specification requires a stable skill domain and therefore this recommendation is critically dependent on the domain specification of Recommendation 1.

We distinguish four decision processes. The first, discussed above, is the specification of a comprehensive and stable domain of skills and knowledge. A second is the selection from that domain of those learning outcomes to be included in a particular assessment. The third is the specification of how these domain components will be grouped for the calculation of subdomain scores. A fourth decision process, also referred to under Recommendation 1, is the specification of the relationship between NAEP exercises and the learning outcomes to be assessed. We are recommending that each of these decision processes be formalized and articulated.

Recommendation 9

NAEP should extend the systematic reporting of distributions of achievement, as well as average levels. The impact of changes over assessment cycles in society, in schooling, and in NAEP procedures on these distributions should be routinely evaluated.

NAEP reporting has been concentrated on averages or central tendency. Some aspects of the Nation's educational attainment are better informed by an examination of the entire distribution of scores. For example, changes over time in average scores may have quite different implications depending upon whether all or only a part of the score distribution is changing. To assist in the examination of changes in distribution, NAEP should report trends for important quantiles of the score distribution. Further consideration should also be given to other methods of representing changes throughout the score distribution.

To assist in interpreting changes in score distributions, it is desirable to isolate the demographic subgroups that contribute to changes in distributions. This may require collection of additional information on schools, students, and variations in administrative conditions. The 1988 report by Beaton, et al. on the 1986 reading anomaly includes a partitioning analysis that reveals the relative contributions to score declines of changes in the mixture of student subgroups and of changes in performance by particular subgroups. This partitioning analysis focuses exclusively on changes in mean performance levels. Such an analysis should be done on a routine basis, and should be focused on the full distribution, not just the means.

Recommendation 10

The expansion of NAEP to provide data at the level of individual States will entail careful study of methods for making and reporting State comparisons. In the 1990 and 1992 pilot studies, a variety of methods should be explored and reported.

Where feasible, State results should be reported for major process and content categories, using the same proficiency scales as are used for National-NAEP. In many content areas, age-specific proficiency scales may be more useful and appropriate than scales spanning different age/grade levels. In addition to reporting absolute levels of achievement on these

scales, each State's performance might be referenced to that of a small group of comparable States, or to nationally representative samples of students matched to State population characteristics. Additional alternatives may also be explored and reported.

Recommendation 11

The reporting of cross-sectional and trend results for State-level NAEP should characterize both the level and distributions of student attainment within each State. This reporting should include (a) demographic subgroup and community differences, (b) variation in performance across major domains of learning outcomes, and (c) distributions of school-level performance within the State.

As discussed in Recommendations 8 and 9, reporting score distributions for major subdomains is more informative than reporting means for broad content areas. This is true at the State level as well as the national level. State and national score distributions for major subdomains should be reported in ways that facilitate their direct comparison to one another.

In addition to distributions for entire States, performance should be reported for demographic subgroups and types of communities within States, whenever such reporting is feasible. Feasibility may be limited by smaller sample sizes for groups or areas within States, or by the legal requirement that results not be reported for schools or districts in the 1990 and 1992 pilot assessments.

Because the school is an important locus of educational policy, we recommend that distributions of school means, as well as distributions of individual scores, be reported. Where samples of schools are sufficiently large and representative, distributions of school means should be reported for States, and for different types of schools within States. By law, particular schools would not be identified.

Recommendation 12

Evaluation of NAEP results, and in particular the sources and magnitudes of errors in estimated achievement distributions, should be undertaken routinely, and not just in response to anomalous findings. Funding should be assured for an ongoing NAEP evaluation, in some way independent of the conduct of the assessment.

To achieve its goals, NAEP should contain a strong evaluation component. This should involve experiments embedded in NAEP that would provide a basis for empirically resolving outstanding issues. The evaluation should include a program of basic and applied research to identify sources of error and model relationships among them. NCES should report NAEP errors on a regular basis, rather than limiting their investigation to apparent anomalies. In conducting the evaluations, particular attention should be paid to (a) NAEP's sensitivity to alternative administration procedures, (b) the impact of saliency of assessment results on individual student performance, (c) the consistency of NAEP results with other measures of achievement, (d) methods

of norming NAEP to relate it to actual performance, (e) the influence of curricular decisions on NAEP outcomes with particular attention to the problem of "teaching to the test," and (f) year-to-year consistency.

Conclusion 1

The anomalous declines in the estimated reading abilities of 9- and 17-year-olds found by NAEP between 1984 and 1986 are much larger than improvements or declines over comparable past periods. After careful study of available evidence, the panel was not able to identify the particular reasons for the reported drop in NAEP reading scores. However, we believe that the most likely primary causes of declines so large and so rapid are changes in testing procedures, and that the 1986 assessment results do not accurately reflect declines in the real abilities of American pupils. Real declines in reading ability may have occurred, but their magnitudes are likely obscured by factors which do not validly reflect changes in pupil learning. The primary causes of the observed decline are still unclear, although we believe that they are probably located in modifications of assessment procedures between 1984 and 1986. New studies currently being conducted by NAEP should help clarify the extent to which such procedural changes were responsible.

Conclusion 2

We believe that differences in college entrance examinations versus NAEP in (1) the populations represented by those tested and (2) the content tested are large enough so that reported trends in college entrance examination scores are not directly comparable to those from NAEP. NAEP was established to serve as the most accurate barometer of achievement for America's young people. Despite its imperfections, we believe that it has and will continue to serve this function better than college entrance examination scores.

Qualification of Endorsement of Recommendations

By Jeanne S. Chall

Comments Regarding Recommendation 2

I dissent from Recommendation 2 because it places almost total emphasis on testing higher level learning outcomes. If NAEP focuses primarily on higher level learnings, the influence of the "lower" on the "higher" learnings will be difficult to determine.

For understanding the course of development in learning such important skills as reading, it would be helpful to assess carefully and specifically the "lower" and "middle," as well as the "higher level" learnings, and also the school, home, and community conditions that affect them.

Statement on Conclusion 1

I dissent with conclusion #1 for the following reasons:

1. If the 1986 reading scores stem from anomalies in testing procedures, one might expect similar declines across all ages tested--9, 13 and 17. Yet while the 9- and 17-year-olds dropped considerably in 1986 over comparable past periods, the 13-year-olds did not.

If the "most likely primary causes" for the 1986 decline are the testing procedures, then it would be necessary to show how the "changes" testing procedures in 1986 affected only the 9- and 17-year-olds and not the 13-year-olds.

2. At the present state of the inquiry into the 1986 reading score declines, I find it difficult to agree with the following statement in Conclusion 1: "the 1986 assessment results do not accurately reflect declines in the real abilities of American pupils" (memo of May 7). As far as I know, no analyses have so far been undertaken to warrant such a statement. While some of the decline may have resulted from changes in testing procedures, it is premature at this point to say that the scores do not accurately reflect the real abilities of American pupils.

Since the possibility of a real decline was considered by only one member of the committee, the following statement is also questionable: Real declines may have occurred, but their magnitude are likely obscured by factors which do not validly reflect changes in pupil learning" (memo of May 7). It would be more reasonable to state that procedural effects were found but that they do not rule out changes that may have taken place in pupil learning.

See in this connection, "Could the Decline Be Real?," the individual report for the subcommittee on the 1986 reading score declines. Based on analyses of NAEP reading trends, it was hypothesized that the increases and decreases in NAEP reading scores from 1970 to 1980, from 1980 to 1984, and from 1984 to 1986 could be explained, in part, by the strengthening and weakening of reading instruction provided by schools, particularly in the early grades, by remedial support when needed, and by support from home and community. When school instruction and support are provided, the scores rise for 9-year-olds (as they did from 1970 to 1980), and they tend to hold up when the same students are tested at ages 13 and 17. When instruction is not as strong in the early grades (as for those tested in 1984), the scores tend to decline and will probably remain low when students reach ages 13 and 17, unless additional measures are undertaken.

The decline in the 1986 reading scores as compared to the 1984 scores tends to follow these trends. While changes in testing procedures may have resulted in the large declines, the possibility of a real decline cannot be dismissed since the 1986 reading scores follow similar trends as those for 1980 and 1984.

For a fuller explication of the hypothesis that the 1986 declines may be real, see "Could the Decline Be Real? Recent Trends in Reading Instruction and Support in the U.S.," paper prepared for the Subcommittee on the 1986 Reading Data of the NAEP Technical Review Panel. See also "Literacy: Trends and Explanations," Educational Researcher, November 1983, pp. 3-8; "New Reading Trends: The NAEP Report Card," Curriculum Review, March/April 1986, pp. 42-44; and "School and Teacher Factors and the NAEP Reading Assessment," paper commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report, The Nation's Report Card, August, 1986 (ERIC Document ED 279 667).

Part 2. Report of the Panel

Edward Haertel, Chair

Report of the Panel

Edward Haertel, Chair

Introduction

The National Assessment of Educational Progress (NAEP) is the only regularly conducted national survey of achievement at the elementary, middle, and high school levels. For the past two decades, it has provided periodic assessments of student proficiencies in reading, writing, mathematics, science, and social studies, and less frequently, citizenship, computer literacy, history, literature, art, music, and career development. In addition to charting patterns and trends in student achievement, NAEP has collected background information that has helped to chronicle changes in educational conditions and practices. NAEP data have been provided freely to researchers interested in conducting secondary analyses, and have supported studies in curriculum, educational policy, methodological research, and research on educational productivity.

As discussed in Dr. Walberg's background paper, the purposes and methods of the National Assessment have evolved over time, and may change even more dramatically in the future. NAEP has evolved in response to new needs and purposes, and to capitalize on new methodologies for data analysis and reporting. With the signing of the Hawkins-Stafford law and the advent of State-level reporting and comparisons, more is expected of NAEP today than ever before. The next several years will bring significant changes in the assessment.

Through its consideration of the 1986 reading anomaly, the accuracy of NAEP trend data, and the challenges that will arise in expanding NAEP to provide State-by-State results and comparisons, the Technical Review Panel has come more than ever to regard the National Assessment of Educational Progress as an invaluable national resource. Problems and deficiencies in NAEP have been identified, but these can be remedied. As it enters its third decade, NAEP is positioned to serve better than ever before as "The Nation's Report Card."

Findings Regarding the Reading Anomaly

Introduction

One of the most frequently assessed and carefully attended areas assessed by NAEP is reading. The reading abilities of 9-, 13-, and 17-year-olds were assessed in 1971, 1975, 1980, 1984, and most recently in 1986. Findings from the most recent reading assessment appeared strikingly different from those of earlier assessments. As stated in the NAEP report Who Reads Best (Applebee, Langer, & Mullis, 1988, pp. 56-57), "The results of the 1986 reading assessment seemed to be out of line with previous NAEP reading assessment results. In particular, they indicated precipitous declines in average reading proficiency at ages 17 and 9. The nature of these drops across only a 2-year period, taken in the context of only modest changes in reading proficiency across a succession of 4-year periods since 1971, was simply not believable."

Declines in scores at both age 9 and age 17 were pervasive, affecting both boys and girls in all geographic regions, racial and ethnic groups, and types of communities. The magnitudes of declines were slightly greater among traditionally lower-performing subgroups, including blacks and Hispanics, children in disadvantaged urban areas, children whose parents have less education, and children who were themselves below the modal grade level. Scores declined more in the southeastern and western regions of the country than in the northeast and central regions.

The absolute magnitudes of the declines at ages 9 and 17 were quite small. Declines in scaled score values were about 3 percent of the 1984 values, reflecting declines in the overall percent correct on reading items of about 3.6 percent for 9-year-olds and 3.3 percent for 17-year-olds. There was a slight improvement in the average scores of 13-year-olds, and there were no concomitant changes from earlier assessments in the 1986 science or mathematics assessments, which were conducted concurrently with reading.

At the same time as the reading means declined at ages 9 and 17, there were striking increases in the dispersion of scores at all three age/grade levels. The standard deviation of reading proficiency scores for 9- and 13-year-olds increased about 10 percent over 1984 values, and at age 17 the increase was about 25 percent. At all three grade levels, the proportions of students at both the highest and the lowest score levels increased from 1984 to 1986. As Hedges observes in his paper, an adequate explanation for the anomaly must explain both the changes in means and the changes in variability.

We invited staff of the Educational Testing Service (ETS) who are working on NAEP, to meet with us and discuss the anomaly; we reviewed the ETS technical report on their investigations of the anomaly (Beaton, Ferris, Johnson, Johnson, Mislevy, & Zwick, 1988); and we examined additional materials, including the actual reading exercises on which the declines occurred, detailed statistical tables not included in the Beaton, et al. report, and trends derived from other data sources.

Two of the panel's background papers, by Hedges and Chall, focused on the reading anomaly. Hedges has provided a technical critique of the ETS report on the anomaly. Chall presents arguments that real declines may well have occurred due to changes in reading curriculum and instruction. The background paper by Wiley also provides analyses that helped to inform the panel's conclusions, and Schmidt's paper touches on related concerns. Walberg's observations should also be noted, that the reading anomaly hardly implies that national and State assessments are unmanageable. Too many well-intentioned procedural changes may have been made too quickly, but as pointed out by Chall and others, actual declines have certainly not been ruled out. More systematic consideration of even apparently minor procedural changes should make future anomalies much less likely, so that unusual performance changes can be more confidently attributed to real changes in ability.

The Panel's Conclusions Concerning the Anomaly

As described in Wiley's paper in this volume, the possible explanations for distributional changes fall into three categories: (a) methodological artifacts, (b) changes in population (e.g., increases in the relative sizes of traditionally low-scoring groups), and (c) changes in student learning. Population changes occur slowly, and would be unlikely to lead to substantial changes over just 2 years. In any case, Beaton's analysis of declines by subgroup (Beaton, et al., 1988, Chap. 6) appears to rule out population changes as the cause of the 1986 score declines. Such rapid and dramatic changes in student learning--either in school or out of school--also appear quite unlikely in the absence of dramatic, simultaneous program changes in schools across the Nation. School curricula and instructional practices do evolve, but such changes are usually gradual, and seem unlikely to have resulted in massive score declines over just 2 years. This analysis leaves methodological artifacts as the most likely primary cause of the observed declines. Specific methodological hypotheses are discussed below.

We were unable to determine the cause of the 1986 reading score declines from the available evidence, but, with the exception of Dr. Chall, we agreed that the most likely primary causes were procedural, and that, although real declines in reading ability may have occurred, their magnitudes are likely obscured by the effects of changes in testing procedures. Dr. Chall explains her reasons for dissenting from this conclusion in her statement following the Panel's Recommendations and Conclusions in this volume, and in her paper. Her argument is also summarized below, in the section on Evidence for an Actual Decline in Reading Scores.

Technical and Procedural Explanations for the 1986 Reading Anomaly

In response to the 1986 reading anomaly, ETS systematically developed and examined a number of possible explanations. These investigations and their results are described in The NAEP 1985-86 Reading Anomaly: A Technical Report by Beaton, et al. (1988). In his paper, Larry Hedges undertook a systematic technical critique of the ETS report, and raised

four general criticisms. First, the report focuses almost exclusively on changes in means, largely ignoring changes from 1984 to 1986 in the shape and variability of achievement distributions at all three age levels. Second, the ETS investigation was organized around the idea that the 1986 reading declines resulted from some single effect, giving little attention to the fact that in combination the different effects considered might well account for the observed declines. Third, the 1984 results were taken as a valid baseline against which to judge the magnitude of declines, largely ignoring the possibility that 1984 results were inflated, and focusing attention almost exclusively on the 1986 assessment as the locus of possible problems. Finally, the ETS analyses may have overstated the precision of the NAEP results. A more complete accounting of sources of error in both the 1984 and 1986 results might have made the 1986 declines appear less remarkable.

Changes in the variance and shape of achievement distributions, as well as means. Changes in the variability in reading achievement scores at all three grade levels were at least as striking as changes in the means. The standard deviation of reading proficiency scores for 9- and 13-year-olds increased by about 10 percent and for 17-year-olds the increase was about 25 percent. The proportions of very high-scoring pupils at all three age levels were actually slightly larger in 1986 than in 1984, but the proportions of very low-scoring pupils were considerably larger among 9- and 17-year-olds. The changes were more complex than simple shifts in means and increases in dispersions. The upper tails of score distributions in 1984 and 1986 are quite similar, but the lower tails of the 1986 distributions are heavier, suggesting declines among some of the students who had been scoring about average.

In his paper in this volume, Wiley presents tabulations made from unpublished ETS data of the score levels corresponding to a series of percentiles in 1984 and again in 1986, for each of the three age groups. He finds that at sufficiently low percentile ranks, there were declines at all three ages, and at sufficiently high percentile ranks, there were increases. The "crossover" point at which 1984 and 1986 scores were the same was at roughly the 78th percentile for 9-year-olds and the 75th percentile for 17-year-olds. For 13-year-olds, the crossover was below the 10th percentile. Thus, the median scores declined at ages 9 and 17, and increased at age 13, reflecting the pattern shown by the means. Wiley's tabulations highlight distributional changes from 1984 to 1986 that are common to all three age/grade groups, and reinforce the importance of attending to changes in score shapes, as well as means of the score distributions.

Separate hypotheses that together could account for anomalous declines. The hypotheses ETS was best equipped to investigate concerned their own procedures for processing the data after they were collected. Failures of quality control and artifacts of scaling were thoroughly investigated, and we concur in the conclusion of ETS personnel that these types of problems are very unlikely to have caused the 1986 reading declines. The investigations undertaken by ETS also appear to largely rule out gross problems in sampling or weighting. There are, however, several

classes of explanations that merit closer attention, as discussed in Hedges's paper.

By and large, 9-year-olds were assessed slightly earlier in the school year in 1986 than in 1984, a difference amounting to an average of 22 days. Seventeen-year-olds were also assessed earlier (18 days) and 13-year-olds were assessed slightly later than in 1984 (4 days). Given that time in school is probably more important than chronological age in determining reading performance, and given that achievement growth is probably nonlinear over the course of the school year, Beaton, et al. may have substantially underestimated the possible magnitude of the time of testing effect, especially among 9-year-olds.

The hypothesis that declines reflected administration difficulties in some but not all schools was investigated by Beaton, et al., but more systematic investigation would be desirable. A search for outliers in the distribution of school means suggested that the anomaly could not be accounted for by difficulties at just a few isolated sites, but more pervasive problems related to overall management of the data collection, especially increases in the size of testing sessions for 17-year-olds, were not thoroughly investigated. At age 17, the average size of the groups tested increased from 20 in 1984 to 35 in 1986, and the maximum permissible session size was increased from 200 in 1984 to 250 in 1986. A comparison of variance components at the school versus individual level in 1984 and in 1986 might have been more informative concerning the overall effects of these changes. If the use of large testing sessions (often at the insisterce of school personnel) or disruptions of these sessions were correlated with student background, this effect might have led to the observed pattern of declines for different student subgroups, as well as the overall increase in score variability at age 17.

Changes in administration conditions, including the design of exercise booklets, the mix of different content areas assessed, the timing of exercise blocks, the sequence of activities carried out in testing sessions, and the size of the testing sessions, may have accounted for a substantial portion of the test score declines. Hedges notes in particular that the 1986 assessment used a "fill in the oval" format for responses which were then machine scored, while the 1984 assessment used a "circle the letter" format for responses which were then entered via keyboard. Any effect of this change would be expected to operate at age 13 as well as ages 9 and 17, but other effects may have served to increase 13-year-olds' scores, masking negative effects at that age level. Difficulties with "fill in the oval" may have been greatest for younger children and for traditionally low-performing subgroups, contributing to the increased variance of 9-year-olds' scores, as well as the observed patterns of score declines across subgroups.

In testing sessions for 9-year-olds, the initial block of background questions was increased from 6 minutes of testing time in 1984 to 15 minutes in 1986, and according to a memorandum from WESTAT (Beaton, et al., 1988, Appendix C), "field staff reported that many of the 9-year-olds were frustrated by the amount of time spent on the block of background questions

. . . Perhaps their concentration or motivation for the reading section was affected." Frustration with the task of providing background information may have been greater among traditionally low-performing subgroups, which would have helped to account for both the pattern of score declines across groups and the increased variance of scores in 1986.

Taken together, if these and other effects are approximately additive, several of them could jointly explain most or all of the anomaly. Hedges summarizes these effects in his Table 1. Hedges also considers the possible magnitudes of these effects for 13-year-olds, and finds that the effects of date of assessment, changing patterns of nonresponse, and artifacts associated with scaling all would have tended to increase scores of 13-year-olds while contributing to declines at ages 9 and 17.

Possibility of positive bias in 1984 assessment results. By focusing almost exclusively on the 1986 assessment, Beaton, et al. (1988) may have overlooked factors that led to inflated score estimates in 1984. Any such inflation would, of course, magnify the apparent decline in 1986. In 1984, 9- and 13-year-olds were assessed on reading and writing together. The WESTAT memorandum (Beaton, et al., 1988, Appendix C) mentions that both reading and writing exercises in 1984 were self-paced, and there were some reports that children spent more than the allotted time on the reading items and less than the allotted time on writing. In 1986, reading exercises were again self-paced, but were administered concurrently with mathematics and science exercises paced by tape recorder. Thus, children were constrained to spend no more than the allotted time on reading. At age 17, Hedges notes that 1984 scores were higher than the trend of earlier assessments would have indicated. If the 1984 results were simply extrapolated from the 1971-1980 linear trend, more than 25 percent of the 1986 anomaly would disappear.

The 1984 assessment results for 17-year-olds may have been inflated if, in response to the widespread adoption of minimum competency testing requirements and the general increase in academic rigor in the early 1980s, there was a temporary increase in the dropout rate, leaving fewer low-scoring 17-year-olds in school. Such an effect would also have tended to reduce the variance of the score distribution in 1984, exaggerating the apparent increase in score variability in 1986.

Accuracy of NAEP achievement estimates. The 1984 to 1986 declines in reading scores at ages 9 and 17 are clearly much too large to be explained by the statistical sampling of respondents or by measurement error. However, such random effects may have contributed to the apparent decline, and it is important to estimate their probable magnitudes. Hedges notes that the standard errors estimated and reported in the Beaton, et al. (1988) are cross-sectional, and do not reflect all of the sources of random error that may have contributed to the apparent score declines.

Evidence for an Actual Decline in Reading Scores

A balanced consideration of the 1986 AEP reading score declines requires consideration of the likelihood that some or even all of the

observed score declines at ages 9 and 17 are the result of real declines in the reading abilities of American school children. If logical or empirical support for real achievement declines exists, it may be that the absence of a score decline at age 13, rather than the presence of declines at ages 9 and 17, constitutes the anomaly. In her paper, Dr. Jeanne Chall refines and extends her analyses of trends through earlier NAEP reading assessments (Chall, 1983, 1986a, 1986b) and presents arguments in support of this position.

Before turning to these arguments, it should be noted that the panel as a whole did not take issue with Dr. Chall's arguments. Her paper presents what may be plausible arguments for the direction of changes in pupil abilities from 1984 to 1986, but in the judgment of the majority of panel members, the magnitudes of these declines over a period of just 2 years are larger than would be expected from changes in curriculum and instruction alone. For that reason, most panel members concur that the primary causes of the declines are probably located in modifications of assessment procedures between 1984 and 1986.

Performance trends may be expressed on the NAEP reading proficiency scale, with an effective range of about 100 to 400. Up to 1984, the largest gains between successive assessments, expressed as points per year on the reading scale, were +.9, +.5, and +.8 point at ages 9, 13, and 17, respectively. The largest declines to 1984 were -.5, -.1, and -.06 points per year. In contrast, the annualized changes from 1984 to 1986 were -2.8, +1.2, and -5.7 points at these three age levels. Conclusion 1, which was endorsed by the remaining panel members, explicitly acknowledges that real declines may have occurred, but holds that their magnitudes are likely obscured by factors related to changes in the booklet design, administration procedures, or related factors.

Influence of early reading instruction. Children learning to read pass through a series of different stages. At different stages, they profit most from different kinds of informal experience and formal instruction. In particular, a too-early school emphasis on comprehension and inference, before children have acquired sufficient skill in phonics and other fundamentals, may be of little value. If such instruction takes teaching time away from word recognition, phonics, and the reading of stories and other connected texts, it may even be detrimental.

Chall argues that beginning in the late 1960s and continuing through the 1970s, beginning reading programs were stronger than before or since. The 1970s were a time of earlier formal instruction in reading, more challenging reading textbooks grade for grade, earlier and more systematic teaching of phonics, Sesame Street and the Electric Company, more remedial help for those who needed it, and Head Start. Since that time, funding for remedial reading instruction has declined, and Dr. Chall contends that a misplaced emphasis on comprehension and inference a early grade levels ("meaning-emphasis") has led to less time for connected reading--and to declining scores. She also cites evidence for the continued importance of early reading instruction to later reading performance.

Patterns of findings from earlier NAEP reading assessments are consistent with the hypothesis that (1) children who were in the primary grades from the late 1960s through the 1970s profited from improved beginning reading instruction; and (2) as these children moved through higher grade levels, they maintained their early advantage relative to other age cohorts. Children in first grade in 1968, 1972, and 1977 were tested as fourth graders in the 1971, 1975, and 1980 NAEP assessments, which showed steady improvement over time. First graders in 1981, tested as fourth graders in the 1984 assessment, did not do as well as those tested 4 years before. Children in first grade in 1964, 1968, 1973, and 1977, tested as eighth graders in the first four reading assessments, showed a slight improvement from 1971 to 1975, a larger improvement from 1975 to 1980, then virtual no change from 1980 to 1984. Among 17-year-olds, performance was essentially flat from 1971 through 1980, then improved from 1980 to 1984.

Dr. Chall extrapolated these trends to predict a continued decline at age 9, stable performance at age 13, and improvement at age 17. At age 9, the direction of changes in both the mean and the variance are in accord with her predictions. According to Dr. Chall, a meaning-emphasis approach to beginning reading (as opposed to a code-emphasis program) would have resulted in general declines, and would have been most detrimental to lower ability youngsters. The 13-year-olds were in the first and second grades in the late 1970s. Dr. Chall suggests that these years were "characterized by a stronger emphasis on word recognition and phonics," and goes on to argue that because they benefited from these early code-based programs, the 1986 13-year-olds "were more prepared to benefit from the emphasis on reading comprehension that they may have received when they were in the intermediate and upper elementary grades."

Concerning the 17-year-old test score decline, Dr. Chall acknowledges that an explanation based mainly on the beneficial effects of stronger beginnings does not seem to hold. Since the 1986 17-year-old cohort was in the primary grades during the 1970s, they would have been expected on that basis alone to show gains. In considering other factors that may have brought about actual declines for 17-year-olds, Dr. Chall proposes as one possibility the publication and wide influence of A Nation at Risk (National Commission on Excellence in Education, 1983) and other "reform reports," published around 1983 and 1984. She observes that these reports called for raising standards and curriculum requirements, increasing the difficulty level of textbooks, and placing more emphasis on higher mental processes. Although some of these reports suggested remedial instruction for the lowest achievers, Chall questions whether much was provided. If changes in the direction of higher standards and more difficult textbooks were implemented, they may have been detrimental for students having difficulty meeting even the lower standards, unless these students received remedial instruction. Dr. Chall goes on to suggest several out-of-school factors that may also have contributed to the sharp 2-year score decline among 17-year-olds. The direction of changes in both the mean and the variance of the 17-year-old score distributions from 1984 to 1986 are in conformity with her explanations.

Efforts by ETS to Resolve the Question of the 1986 Reading Anomaly

As part of the 1988 NAEP data collection, ETS is conducting parallel data collections replicating as closely as possible the procedures followed in the 1984 reading assessment, and also the procedures followed in 1986. If 1986 reading scores were influenced by one or more of the changes in testing procedures introduced in 1986, and if these changes act in the same way to influence respondents in 1988 as they did in 1986, then comparing the results of the 1984 replication and the 1986 replication will yield an estimate of the adjustment that must be made to the original 1986 results to make them comparable to results from earlier assessments. As stated in the discussion of Recommendation 5, the panel strongly endorses these efforts as a vital source of information for future design decisions. At the same time, we note that ad hoc investigations of apparent anomalies do not provide adequate quality control to assure the reliability and validity of inferences from NAEP about trends in achievement.

It is unfortunate that the results of the procedural comparisons being carried out as part of the 1988 assessment are not scheduled to be made available until sometime in 1989. One initial reaction is that a small, quick study should be mounted to compare the 1984 and 1986 procedures and get some answers. However, given that the effect to be detected amounts to a change of only a few percent in item difficulties, and given that replication implies the use of BIB spiralled booklets, a large, systematic study may be the only way to get satisfactorily definitive answers.

Findings Regarding Trends

Introduction

NAEP was established to provide accurate information at the national level about school achievement, including changes over time. Indeed, the title "National Assessment of Educational Progress" expresses that intent. In their 1987 Study Group report, The Nation's Report Card, Alexander and James reaffirmed this historic commitment, and this present Technical Review Panel also concurs strongly in the priority for NAEP of providing accurate and trustworthy information about achievement trends.

Even though NAEP was explicitly designed to provide accurate longitudinal information concerning student achievement, the public, educational policymakers, and even scholars have often relied on other information sources when drawing conclusions about achievement trends. The widely publicized test score decline from the middle of the 1960s until the beginning of the 1980s was more generally associated with the SAT than with NAEP, and the question often arises whether the SAT, with its much larger and perhaps better motivated samples of examinees tested annually, provides better trend information for some purposes than NAEP, with its small samples, biannual schedule, and assurances of student anonymity.

Concerns over the accuracy of NAEP data for charting trends and making comparisons have been heightened by the 1986 NAEP reading anomaly, and also by the increasing interest in State-level achievement comparisons. The Secretary of Education's "wall chart" presently uses SAT and ACT scores for State-level achievement comparisons, and under the Hawkins-Stafford law, the 1990 and 1992 NAEP assessments will report achievement and achievement comparisons for participating States at selected grade levels in mathematics (1990 and 1992) and reading (1992).

In the light of these concerns, this panel was charged with addressing the question of whether NAEP was the most accurate barometer of trends in the achievement of American school children. Four of the panel's background papers, by Baldwin, Pandey, Wiley, and Schmidt, specifically addressed issues of achievement trends. Dr. Baldwin's paper focuses on trends in reading, and analyzes changes in NAEP reading objectives and in analysis and reporting procedures across the five reading assessments from 1971 through 1986. Dr. Pandey's paper addresses changes in the framework of NAEP mathematics objectives across the four mathematics assessments from 1973 through 1986, and also presents systematic comparisons between trends reported from NAEP and trends derived from other sources. In Dr. Wiley's paper, he systematically compares the content of the SAT verbal subtests and the NAEP reading exercises, compares the populations represented by NAEP versus SAT examinee samples, and provides tentative comparisons between 1984 to 1986 changes in SAT scores and 17-year-old NAEP reading results for the highest achieving examinees. Dr. Schmidt's paper reviews a range of inconsistencies in NAEP procedures that may have compromised trend reporting, and calls for a more systematic procedure for considering any changes from one assessment cycle to the next.

The Panel's Conclusions Concerning NAEP Trends

The panel concluded that in comparison to other national, longitudinal data sources, NAEP provided the most accurate and useful information available. At the same time, we found significant deficiencies in NAEP, and have recommended several changes to improve the quality of NAEP trend data for the future. As stated in our Conclusion 2, college entrance examinations and NAEP sample different examinee populations and different domains of content. These differences virtually preclude valid comparisons between reported trends from NAEP versus college entrance tests, although it is possible that valid comparisons might be constructed for higher-ability 17-year-olds through careful analyses of portions of the data from these two sources. Although some types of items appear on the SAT and not in NAEP (e.g., vocabulary items), we believe that NAEP is the better barometer of national trends. It more faithfully represents the entire school population, tests younger as well as older children, samples a broader range of content areas and of objectives within content areas, and provides more detailed score reporting. Moreover, NAEP is designed specifically to assure the continuity of trend lines. In contrast, SAT and ACT trends are merely by-products of examinations designed for a very different purpose.

That being said, NAEP trend reporting nonetheless presents some difficulties. The panel noted that in both reading and mathematics, there have been substantial inconsistencies over the years in the content of NAEP exercises, in the frameworks of objectives used to organize those objectives, and in forms of reporting. Successive assessments within a given content area (e.g., reading or mathematics) have been linked using common items, but the different sets of linking items used over the years have sometimes represented quite different mixes of subdomains. (Such subdomains include, among others, Literal Comprehension, Inferential Comprehension, and Reference Skills in reading; or Numbers and Operations--Knowledge, Fundamental Methods, and Measurement in Mathematics.) Our concern with these issues is reflected especially in Recommendations 1 and 5.

In addition to noting inconsistencies in the content of successive assessments, the panel was also concerned about the overall number, scope, and quality of NAEP exercises. Different panel members noted an overreliance on exercise formats calling for selection rather than production of correct responses, and insufficient coverage of higher-level learning outcomes, as well as failure to measure and distinguish important fundamental skills usually prerequisite to higher learnings. In Dr. Pandey's paper, he observes that most free-response exercises in mathematics looked like multiple-choice questions with the answers removed. These concerns are expressed in our Recommendations 2 and 8.

Comparison of Trends Reported from NAEP Versus Other Data Sources

Before considering the relative accuracy of achievement trends revealed by NAEP versus other data sources, it is well to consider the extent to which they agree. Pandey's paper presents comparisons of NAEP with

other data sources in mathematics, and Wiley's paper discusses comparisons of reading trends from NAEP versus the SAT.

Pandey's comparison of mathematics achievement trends from different data sources indicates substantial agreement in the directions of performance changes over time, although the relative magnitudes of changes shown by different data sources are difficult to compare directly. Comparability is limited by the definition of NAEP samples prior to 1983 according to age rather than grade level; and by the fact that alternative longitudinal data sources generally are not nationally representative. Other differences noted by Pandey include the time of year of testing, test administration procedures, and the content of different tests. Despite these limitations, NAEP trends were compared to data sources including the SAT, American College Testing (ACT) program, General Educational Development (GED) examination program, National Longitudinal Study (NLS) of the High School Class of 1972, High School and Beyond (HSB) study, and the Iowa Tests of Educational Development (ITED). The data summaries and tabulations made by Koretz (1986, 1987) were used extensively in making these comparisons. Within the limitations of the data, trends revealed by NAEP appear consistent with those derived from other data sources.

Wiley's paper offers a more detailed analysis of the limitations of comparisons between NAEP and SAT reading trends, focusing especially on 1984 to 1986 performance changes among 17-year-olds. The principal limitations on such comparisons are differences in the populations sampled and in the kinds of items included. Differences between populations can be accounted for if one assumes that above some ability level, virtually all examinees would have taken the SAT. Under this assumption, and using data on the proportion of all in-school 17-year-olds taking the SAT, it is possible to derive corresponding percentile ranks in the SAT and NAEP achievement distributions. This amounts to comparing the most able students in the NAEP sample with the most able students in the SAT sample. The calculations required are presented in Wiley's paper.

Content differences can be reduced but not eliminated by using only the reading comprehension subscale of the SAT verbal scale (i.e., excluding the vocabulary subscale). Compared to SAT reading comprehension items, the NAEP reading exercises span a broader range of difficulty levels. A relatively small proportion of NAEP exercises are as difficult as those in the SAT. However, these more difficult NAEP exercises are probably the most discriminating for high-ability examinees, and so functionally the tests may measure similar skills at high ability levels. Careful consideration of the difficulties in validly comparing NAEP and SAT trends highlights the limitations of the SAT as a general barometer of educational achievement.

When the procedures developed in Wiley's paper are applied to NAEP and SAT data in 1984 and in 1986, both data sources show improvements from 1984 to 1986. (Recall that the increase in variance of achievement scores led to increases from 1984 to 1986 in the proportions of very high-scoring examinees as well as very low-scoring examinees at all three age levels.) The magnitudes of these examinees' improvements according to the SAT versu.

NAEP probably cannot be validly compared. In any case, no such comparison has been attempted.

Accuracy and Interpretability of NAEP Trends

In both reading and mathematics, the accuracy and interpretability of NAEP trends have been diminished by changes over assessment in the frameworks used to organize objectives, and by changes in the mix of exercises used to link successive assessments. In Baldwin's paper, she tabulates the proportions of Literal Comprehension, Inferential Comprehension, and Reference Skills exercises included in each of the five NAEP reading assessments from 1971 through 1986, and also the numbers of exercises in each category that were common to more than one assessment. She finds that trends in 17-year-olds' achievement differ from one item type to another, and that a different mix of subdomains was used in linking the 1984 and 1986 assessments than had been used to link earlier assessments. In particular, the proportions of Literal and of Inferential Comprehension exercises each dropped by about 10 percent, and the proportion of Reference Skills exercises increased from 15 percent to 34 percent.

These findings have two major implications. First, the meaning of trend comparisons has not been entirely stable over time. Second, exclusive reliance on the NAEP reading scale in examining and interpreting achievement trends might obscure important differences in performance changes for different kinds of reading skills. This is not to say that summaries like that provided by the NAEP reading scale are without value. Such scales can provide readily interpretable summaries of broad trends, and can be useful in guiding educational policy. However, they must be supplemented with more refined scales focused on component skills, as discussed in the panel's Recommendation 8. The unidimensional models on which the reading and other NAEP scales are based are only approximations. Important differences exist among subdomains in reading and other content areas. For that reason, it is important that the mix of exercise types comprised in such scales be held constant over time.

In Pandey's paper, he presents an analysis of NAEP mathematics objectives through time that shows even greater variation over time than in reading. Pandey reclassifies exercises from earlier assessments in terms of the most recent objectives framework, and tabulates the numbers of exercises in each category that were common to two or more of the 1978, 1982, and 1986 assessments. His tables show, for example, that at the 9-year-old level, between 1978 and 1982, 28 percent of the common items were from the categories Numbers and Operations--Knowledge and Numbers and Operations--Applications, and 13 percent were from the category Fundamental Methods. Between 1982 and 1986, the corresponding percentages had changed from 28 percent to 41 percent, and from 13 percent to 7 percent. Less dramatic changes also occurred for other content categories, and at the 13- and 17-year-old levels. These comments should not be taken to imply that the content and meaning of NAEP scales should never change, only that such change should be planful and deliberate, not accidental. An unintended consequence of the consensual process used to arrive at sets of objectives

for each separate assessment may have been diminished attention to consistency through time.

Recommendations for NAEP in 1990 and Beyond

Introduction

The panel's final charge was to consider issues that will arise in the expansion of NAEP to provide State-level achievement comparisons. The Hawkins-Stafford law calls for State-level comparisons on a pilot basis in 1990 and 1992, with States participating on a voluntary basis. In 1990, State-level comparisons will be made in mathematics at a single age/grade level. In 1992, pilot State-level assessments will be conducted in mathematics at two age/grade levels, and in reading at one level. The expectation is that at some point beyond 1992, State-level comparisons will be expanded well beyond these pilot studies.

State-level comparisons have already been made on a regional basis by the Southern Regional Education Board (SREB). With the cooperation and assistance of the Educational Testing Service, participating States conducted assessments modeled after NAEP and compared their performance among States and against national performance levels. In addition to States participating in the SREB comparisons, ETS has provided State-level assessments to several other States, augmenting National-NAEP samples and collecting additional data as part of the regular National-NAEP data collection. These experiences with State-level assessments have helped to highlight technical and administrative issues likely to arise in the anticipated NAEP expansion, but the 1990 and 1992 pilot State assessments will provide far more information. The panel's deliberations, background papers, and recommendations are intended first of all to guide the 1990 and 1992 pilot assessments. Although we believe that our conception of an assessment supporting State-level comparisons is sound, we anticipate that decisions about the shape of the assessment after 1992 will be informed by the results of pilot studies over the next several years.

Design and Administration of National and State-Level NAEP

There is no fundamental difference between the organization and activities of an assessment designed to provide national achievement estimates and one reporting at the level of the separate States. However, differences in the intended uses of the data collected, in concomitant incentives on the part of students, teachers, and test administrators, and in scale and cost have implications for the design of an expanded assessment program.

Even apart from the expansion of NAEP to accommodate State assessments, the panel would call for some changes in the design of the assessment. The 1986 reading anomaly as well as limitations in the validity and interpretability of NAEP trends suggest a need for improvements, even if the scope and purposes of the National Assessment were to remain as they are now. The design recommendations in this section address both the improvement of NAEP at the national level and the expansion of NAEP to provide State-level comparisons.

A first, major question in the expansion of NAEP is whether separate data collections should inform national versus State-level achievement estimates, or whether a single data collection should serve both purposes. Ultimately, it may be that national estimates will be obtained from the union of State-level data collections. For 1990 and 1992, however, it seems clear that a national-level data collection will be designed following essentially the same procedures as in the last two or three assessments, with augmentations to the national sample in States electing to participate in the pilot State-level comparisons. A sample designed to provide good estimates for the Nation would not be the same as one designed to provide good estimates for the separate States. Moreover, abrupt changes in the National-NAEP sampling plan could jeopardize the continuity of national trends. Most importantly, State participation in NAEP will be voluntary in 1990, 1992, and beyond. Valid national estimates could not be assured if NAEP were to rely on data from some arbitrary subset of States electing to participate.

In this section, the terms National-NAEP and State-level NAEP will be used to refer respectively to the National-NAEP data collection (and national-level data analysis and reporting) and to the State-level augmentations (along with State-by-State analysis, reporting, and comparisons). The precision of National-NAEP estimates may be enhanced by using State-level data, but these data cannot be combined in any simple way unless they are collected using instruments that are identical in every important respect, administered under carefully standardized conditions to samples having a known relationship to the National-NAEP sampling frame.

Whether or not State-level NAEP data are used in formulating national achievement estimates, comparability between State and national data collections is critical. A primary purpose of State-level data collections will be to enable comparisons between State and national achievement levels, as well as comparisons among States. Mechanisms to ensure such comparability are addressed most directly in the background papers by Musick and by Bock, although most of the papers touch upon these concerns to a greater or lesser extent. The evaluation and quality control mechanisms set forth in Recommendation 12 and in the paper by Dr. Schmidt would also go a long way toward assuring the dependability of such comparisons.

Recommendations concerning State-level NAEP procedures. The two papers by Musick and by Bock each propose specific procedures for a State-level assessment. Dr. Musick draws on his experience in State-level comparisons with the Southern Regional Education Board, and Dr. Bock draws on his experience in working with several States on designs for assessment programs. The two papers are complementary--Musick addresses primarily issues of administration, governance, and the logistics of data collection. Bock addresses primarily technical issues in the design, analysis, and reporting of assessment results. The proposals expressed in these two papers have evolved through the course of the panel's deliberations. By and large, the panel has reached consensus on at least the broad outlines of a design, the general features of which are sketched below. Justifications and supporting details may be found in the

background papers, especially those just cited.

The panel conceives of State-level NAEP as a program unit under the parent National-NAEP organization. It would have its own staff director and would be supported by its own advisory structure, reflecting State interests and concerns. State-level assessments would be conducted in conjunction with National-NAEP assessments, within the same time frame, which should be shorter than the present 12-week testing schedule for national biannual assessments.

Present NAEP data collection procedures are designed to be minimally intrusive for participating schools. Once a school is contacted, however, the incremental costs of testing more students within that school or of increasing the testing time are relatively small. The more direct involvement of the States in NAEP offers an opportunity to reconsider decisions about testing burden. For a variety of reasons, the panel believes that testing time should be increased. (See Recommendation 4.) A two-stage testing approach might further increase the efficiency of the assessment, helping to assure that the best possible use was made of students' time. Possible methods of implementing this and related procedures are presented in Bock's paper.

The utility of the State-level and National-NAEP can be dramatically increased by providing methods of linking to them the autonomous testing and assessment programs of individual States. States that had suitable testing programs and chose to carry out such linking could then report results of their own testing for schools or districts, in terms of the NAEP scales. In Bock's paper, he describes feasible methods for accomplishing such linkages while assuring the confidentiality of NAEP results for schools and students, as required by law.

Assuring comparability. As stated in our Recommendation 6, it is critical that the administration procedures for State- and National-NAEP be the same in every important respect. However, these need not be the same as present NAEP procedures. For example, the quality of both State and national data could probably be increased if students were tested in groups of no more than 30, and if two adults were present at each testing session. One would be an external examiner trained under the direction of the National-NAEP contractor, and the other would be someone from the local school with whom the students were familiar. This would help prevent disruptions which might depress the scores of entire groups of students, and would help to minimize the "substitute teacher effect," especially with younger children. The external examiners would probably be persons provided by the State for the time required for training and test administration. Personnel might be recruited from the field staffs of large sample survey firms, from the ranks of professional substitute teachers, or from the faculty of community colleges, for example. External examiners would be chosen to minimize travel and overnight lodging expenses, but would not manage any assessments in the school organizations by which they were employed. Note that changes in the size of testing sessions or in the number of adults present would call for systematic review, and for bridge studies to assure the continuity of trends, as

discussed in Recommendation 5.

In addition to administration procedures for State-level and National-NAEP that are the same in all important respects, comparability requires the use of common instrumentation. The occurrence of the 1986 reading anomaly suggests that some items may possibly function differently in the context of different exercise booklets, assumptions of item response theory notwithstanding. To the extent possible, State-level NAEP exercise booklets should be identical to National-NAEP booklets. If the content of State-level NAEP is less comprehensive than that of National-NAEP, then the State-level NAEP booklets should correspond to a subset of the National-NAEP booklets. As mentioned in Recommendation 6, States may choose to supplement the core State-level NAEP data collection, but any supplemental questions should follow the core instrumentation.

Common administration procedures and common instrumentation are two of the four critical requirements for assuring State-level and National-NAEP comparability. The remaining concerns are first, comparable sampling of schools and students within schools; and second, uniform procedures for determining which students should be excused from testing because they are of limited English proficiency, educable mentally retarded, or functionally disabled.

Samples of schools and of students within schools should be drawn under the supervision of the National-NAEP contractor or subcontractor responsible for the National-NAEP sample. The same sampling frame should be used, although of course the selection of numbers and proportions of schools selected within strata may differ. States may assist in the sampling by providing technical or clerical assistance or, if necessary, lists of schools. Selection of students within schools should likewise follow the same procedures for State-level as for National-NAEP.

Exclusion criteria for limited English proficiency, educable mentally retarded, or functionally disabled students must be defined in the same way for State-level and National-NAEP, and must also be applied in a consistent fashion. Ultimately, these criteria will be interpreted by hundreds or thousands of individuals at the local school level. Detailed written procedures and careful training can help to assure an acceptable degree of uniformity, but in addition, individually written justifications for each student excluded and random audits may be helpful in assuring compliance.

As procedures are developed for all aspects of the State-level NAEP data collection, State testing and assessment personnel should be involved. Procedural manuals should be written to provide a common authoritative reference and to help assure compliance.

Cognitive Items

From its inception, NAEP has espoused the goal of measuring the full range of significant learning objectives in different content areas. The NAEP exercises, the tasks set for students to find out what they know or can do, are the heart of the assessment. No refinements in sampling or

administration or statistical methodology can compensate for deficiencies in the scope or quality of the exercise pools.

In the light of recent concerns over the quality of NAEP trends, and in the light of the new purposes that will accompany NAEP's expansion to provide State-level estimates, the quality of NAEP exercises has taken on even greater significance. The panel has serious concerns about the size and scope of the NAEP exercise pools and about the stability over time of their organizing frameworks. Even our consideration of the 1986 reading anomaly was impeded by the impossibility of distinguishing 1984-to-1986 changes with respect to different constructs within the area of reading. Several of the background papers address the quality of the NAEP exercise pools, including the papers by Pandey and Baldwin, but these issues are most extensively considered in the paper by Guthrie and Hutchinson.

Continued attention to fundamentals, and increased emphasis on higher level learning. The advent of State-by-State comparisons will bring NAEP more than ever into the public eye, and will increase pressures to "teach to the test." In itself, this need not be a bad thing. Many States, districts, and schools have consciously used testing to shape curriculum and instruction. But if NAEP's influence on curriculum and instruction is to be salutary and not detrimental, then its exercise pools must be comprehensive. As discussed in our Recommendation 2, the NAEP exercises must embrace both fundamentals and higher level learning. NAEP must assess the intended learning outcomes of typical American school programs, but must also reach beyond the typical to point directions for improvement. Decisions about the learning outcomes assessed should reflect the best thinking of scholars and subject matter specialists.

Testing process learning outcomes in concert. In most content areas, particularly reading, writing, and mathematics, several distinct cognitive processes may be logically distinguished and identified with different kinds of exercises. These may include lower-level and higher-level processes (e.g., word attack versus inferential comprehension) or processes at a comparable level (e.g., inferential comprehension and interpretation). There is a tension between using exercises that call for these processes separately (e.g., exercises to assess word attack skills) versus exercises that call for their use in concert. Careful, scholarly deliberation will be called for to resolve questions about the granularity of both exercises and reporting scales in each content area assessed.

One primary consideration is the intended scoring and reporting of assessment outcomes. Once skills are combined at the level of the NAEP exercises, it becomes difficult if not impossible to report them separately, or to disentangle their relative contributions to possible low performance. It is neither appropriate nor useful for NAEP to aim for detailed, diagnostic profiles of learning strengths and weaknesses. At the same time, the assessment should be sensitive to and able to distinguish broad changes over time in curriculum focus and emphasis, for example, the direct instruction of high-level strategies in reading, or greater problem solving focus in mathematics.

A second consideration is the factorial structure of the processes involved. If important component processes have low intercorrelations, then they should probably be assessed separately, but if they are highly intercorrelated, separate tests and test scores may be redundant. It must be recognized that high intercorrelations among items in an exercise pool may result from instructional practices and curricular organizations, as well as the logical, substantive structure of the exercises themselves, and the inherent nature of children's cognitive development and information processing. However, it is likely that exercises could be organized into two or more independent scales at each age/grade level in reading, mathematics, or writing, acknowledging the complexity of these areas while providing sufficient statistical independence that scores would be separately interpretable.

Consideration of efficiency may argue in favor of more complex, integrative tasks that assess lower-level processes as components of more complex performances, making separate tests of those processes unnecessary. A long division problem may test division, subtraction, and multiplication all at once. At the 13- and 17-year-old levels, and probably as early as the 9-year-old level, reading is a critical tool to support learning in other school subject areas, and mathematics may be important for representing, understanding, and manipulating concepts throughout the curriculum. As a consequence, exercises that assess the ability to use reading and mathematics as "tools" for problem solving in content areas should be included.

Finally, findings from cognitive psychology may also argue in favor of more integrated assessments of process. As discussed in Guthrie's and Hutchinson's paper, logically separable processes develop in concert, and support one another. Full mastery of one process may be impossible without partial mastery of others. Moreover, the ability to apply different processes appears to be context bound. It is notoriously difficult to achieve transfer of learned skills to different applications. If reading processes like monitoring for understanding or using previous knowledge to understand new ideas are tested in isolation, an incentive is created to teach them in isolation, and for many children, their concerted application in actual reading may be far from automatic.

Separation of content and process. The knowledge or skills an exercise is designed to assess may be referred to as its intended requirements. The intent of the exercise is to distinguish those that do or do not possess the attributes it is designed to measure. But every exercise also calls for additional knowledge, skills, and dispositions. At the very least, valid assessment of the intended requirements relies on knowledge about the test-taking situation, skill in marking answers accurately, and a disposition to attempt each exercise seriously. These and other additional skills may be referred to as an exercise's ancillary requirements. Valid assessment of intended knowledge and skills is only possible if an exercise's ancillary requirements are held to a level that can be presumed nearly universal among the group tested.

Content knowledge is often ancillary to the measurement of process, and processes like reading and writing are often ancillary to the measurement of content. Thus, exercises intended to assess content knowledge and not reading ability should be written at a reading level at least 2 years below the grade level of the examinees. Likewise, exercises intended to assess reasoning or problem solving should have children apply these skills to material or situations with which the great majority should be very familiar. An assessment that purports to show the extent of historical knowledge should not be confounded with students' reading ability, although reading in the content area of history might in itself be a worthwhile thing to assess. Likewise, scores on an assessment of critical thinking should not be confounded with content that is known to some students and unknown to others.

Background Items

Exercises measuring learning outcomes may be at the heart of the assessment, but valid and reliable measurement of learning outcomes alone is not enough. NAEP achievement data become useful for policy when they can be related to other variables. Background questionnaires for students, teachers, and principals are administered concurrently with student achievement exercises. These permit the reporting of NAEP results for major subpopulations (e.g., gender and race/ethnicity) and in recent years have also provided limited information on instructional processes, so that these could also be related to schooling outcomes. In the panel's deliberations, some of the same concerns were raised about these background instruments as about the cognitive exercises. The panel saw a need for greater consistency through time in the questions asked, and for a stable and coherent framework to guide the selection of background questions. As stated in Recommendations 3 and 4, we believe that the time allocated to background data collection should be increased, especially in the light of new purposes accompanying the advent of State-by-State comparisons.

Several of the background papers address these concerns, but the panel's positions are developed most extensively in the paper by Baron and Forgione. In their paper, Dr. Baron and Dr. Forgione draw upon their experience with the Connecticut Assessment of Educational Progress, and make extensive use of Dr. Jeannie Oakes's framework for organizing alterable variables in education. Their paper includes appendices giving some useful classifications of variables, although they caution that not all of these are important to assess. Baron and Forgione also call attention to the need for coordination and triangulation of questions on student, teacher, and principal questionnaires, bearing in mind that these classes of respondents bring different perspectives to bear on schooling processes.

NAEP background questions serve a variety of purposes in the assessment. There is no end of interesting questions to ask, and so a judicious, disciplined selection of background questions is essential. The panel proposes that each background question should represent at least one of three broad categories. First would be "unalterable," or demographic variables important for describing patterns in NAEP achievement data.

Second would be variables plausibly related to achievement, including indicators of curriculum content and orientation or of instructional practices. Third would be variables reflective of valued schooling outcomes that cannot be directly captured by NAEP achievement exercises. Each type of variable is described below.

Variables important for describing NAEP achievement patterns. Examples of unalterable values (the first category) are gender, race/ethnicity, socioeconomic status, and size and type of community. These variables are "unalterable," but many educational policies and philosophies are predicated on the well founded assumption that their relations to schooling outcomes are alterable.

Variables plausibly related to achievement. Variables plausibly related to achievement include questions about instructional practices, for example, a "whole language" approach to reading, writing, speaking, and listening. Baron and Forgione report limited success with such questions in their own experience, in part due to the validity of the questions themselves, and in part due to the complexity of the relationships of these variables to learning. For example, teachers may give more feedback on papers to low achieving students, so a simple correlation appears to show that teacher feedback is negatively related to achievement. A more promising focus for questions in this category may be on student opportunity to learn. If both the quality and the quantity of relevant content coverage can be addressed, such background questions may emerge as powerful predictors of learning outcomes. Questions about homework and out-of-school pursuits can also help to inform the sum total of students' educative experiences. Specific examples of background questions in a range of content areas are provided in Baron's and Forgione's paper. Finally, this second category of background questions plausibly related to achievement might include a few concomitant measures of achievement that might be used to validate patterns of NAEP findings, e.g., "What grades do you usually get in school?" (or in some particular content area being assessed).

Variables reflective of other valued schooling outcomes. The third category of variables to be represented among background questions include, for example, amount of leisure reading, or participation in student elections. A range of affective or attitudinal schooling outcomes should also be sampled in the NAEP background questions. In their paper, Baron and Forgione offer by way of illustration two statements from a Connecticut State testing program, with which students were asked to agree or disagree: "Careers in science are more appropriate for men than for women" and "My knowledge of science will be of little value to me in my day-to-day life."

Educational indicators as ends in themselves. Once any indicator is assessed and reported, improvement with respect to that indicator may become an end in itself. This is true of both achievement exercises and background questions. Thus, background questions should be selected such that direct efforts to improve a school's standing with respect to those questions would be salutary for education. If counts of courses taken are reported, for example, there may be an incentive to offer a greater number

of "watered down" courses, with no concomitant improvement in student learning. Following Mehrane (1987), Baron and Forgione argue that this corruptibility of indicators can be diminished if they are specifically defined and include a qualitative as well as a quantitative dimension.

Analysis and Reporting of Results for the Nation and for Participating States

The expansion of NAEP to provide State-by-State comparisons raises a number of new issues in the analysis and reporting of results. Perhaps foremost among these is the problem of reporting interstate comparisons, but State-to-national comparisons, State trends over time, within-State comparisons among regions or student subpopulations, and the reporting of distributions of school means, as well as student-level achievement distributions, all call for attention. A few of these issues may be set aside, for the present, in the light of the Hawkins-Stafford law's prohibition against reporting results for identifiable units below the State level of aggregation. The panel did not give detailed consideration to this entire range of issues, but did consider a number of them.

Recommendations 7, 8, 9, 10, and 11 all address analysis and reporting issues. At both the State and national levels, we recommend that assessment design and analysis permit the accurate estimation of scores for individual students. Specifically, we call for a retreat from the "plausible values" used as the basis for recent NAEP reports. We call for increased reporting of subdomain scores using scales specific to the content appropriate to specific age/grade levels where such scales are more appropriate than scales common to two or more levels. The panel also calls for fuller reporting of score distributions than is provided by means alone. Where feasible, these recommendations should be implemented in parallel fashion at the State and national levels. Finally, the panel calls for systematic study of alternative methods for making and reporting State comparisons. The panel recognizes that taken together, these recommendations imply increases in the amount of data collected (cf. Recommendation 4).

Issues in reporting at both State and national levels. Many of the reporting issues raised by the panel apply equally to National-NAEP and State-level NAEP. In addition to concerns addressed earlier in this Review of Findings, panel members addressed the use of NAEP proficiency scales that span the range from age 9 through age 17 (e.g., the present NAEP scales in reading and in mathematics); the value of reporting distributional summaries at the level of school means as well as score distributions for individuals; the timeliness of NAEP reporting; and the importance of relating NAEP performance to "real-world" schooling outcomes.

The use of common reporting scales across age/grade levels is problematical for at least two reasons. First, the range of such scales must necessarily be so broad that important within-grade variations are obscured because they occur over a narrow range of scale values. Second, such scales are difficult to interpret when they represent qualitatively different kinds of content at different age/grade levels. In mathematics,

for example, the topics taught and tested for 17-year-olds may overlap little with those for 9-year-olds. The fact that an item response theoretic (IRT) model can be applied to data from three age levels combined does not assure that the results will be sensible.

Data analysis and reporting should take cognizance of the hierarchical nature of educational data. Methods for multilevel analysis and reporting should be used to present results for schools as well as for individuals both for States and for the entire nation.

It was argued in several papers that the present 18-month turnaround between data collection and reporting of results will be unacceptable for purposes of State-level comparison. The NAEP contractor should make basic statistical data available as soon as reasonably possible, before interpretative reports are written. At the same time or shortly thereafter, public use data tapes should be made available for secondary analysis. At the same time, as Dr. Burstein observes, raw, unelaborated columns of numbers may be inaccessible to important policy audiences. Concurrent release of data and interpretations is an important means of retaining control over the meanings imputed to the data and the kinds of recommendations they are used to support.

As discussed in Dr. Bock's background paper, the meaningfulness of NAEP reporting scales could be enhanced significantly by empirical studies relating performance on the NAEP scales to more directly measured, practical schooling outcomes. Measurement of students' ability to perform real-world tasks is far more costly than collecting data using paper-and-pencil measures. However, valid inferences about the population's performance on such tasks may be based on large-scale assessment using NAEP exercises, together with much smaller studies to determine the relation between NAEP scales and such real-world outcomes. Such studies would significantly advance the kind of construct validation envisioned by Guthrie and Hutchinson.

Reporting of State-level results and interstate comparisons. State-by-State comparisons are addressed in several of the background papers, but especially the papers by Dr. Burstein and Dr. Haertel. Burstein observes that the panel's consideration of issues in reporting is complementary to recent work by the Council of Chief State School Officers (CCSSO) in modeling the consensus planning process recommended by the Alexander and James Study Group report, The Nation's Report Card. The panel's recommendations and those of the CCSSO should be largely compatible. We concur with the CCSSO that NAEP must measure a range of important learning outcomes, and that the system developed must not merely provide gross, simplistic State comparisons of the kind often seen with comparative school achievement data, but must place achievement patterns in the context of possibly different educational goals, demographics, and other contextual factors. Specific recommendations from this panel versus the CCSSO Consensus Planning Project are contrasted in Dr. Burstein's paper.

Burstein's paper briefly reviews the context and assumptions surrounding the panel's consideration of analysis and reporting issues, and

then turns to the purposes of State-level NAEP reporting. NAEP must provide a reliable and valid assessment, making efficient use of the student time taken for data collection. Reporting must take account of the different needs and circumstances of the several States. To be useful in guiding policy, it should relate achievement to alterable variables--concrete features of the school systems that can be changed for the better by State and local educators. Fair and credible reporting of State-by-State comparisons is essential if State cooperation is to be enlisted and maintained. As stated in Recommendation 10, the 1990 and 1992 pilot assessments should be used to explore several alternative schemes for making and reporting such comparisons. Following in part on work by the CCSSO, Burstein recommends several specific methods that should be explored for making and reporting State-level comparisons.

In Haertel's paper, he considers models used within States for school- or district-level comparisons, and considers their applicability to the problem of State-by-State comparisons. Haertel concludes that it will probably be necessary to place State achievement disparities in the context of broad differences in socioeconomic level, although in principle reporting of unadjusted means for subpopulations within States could suffice. He cautions, however, that patterns of actual achievement must not be obscured. Raw and contextualized reports of State-level achievement differences serve different sets of purposes, both important. Adjustments must not be permitted to legitimate existing inequalities in educational outcomes for different groups of learners.

Evaluation

The investigations triggered by the 1986 reading anomaly have called attention to a serious need for more systematic, ongoing statistical evaluation and audit of NAEP procedures and results. The kind of evaluation referred to would not consider the value or utility of NAEP, but would examine closely the statistical quality of NAEP findings. The papers by Hedges, Schmidt, Bock and Musick touch on several of the panel's concerns. First, there is a need for empirical studies of the error structure of the assessment. Expert judgments, including those of this panel, cannot resolve fundamentally empirical questions. Bridging studies need to mirror the procedures of the main data collections in every important respect.

Second, studies of the accuracy and quality of reported NAEP results need to be conducted on a routine basis, not just in response to apparent anomalies. Third, to the extent possible, statistical evaluation of NAEP should address the full range of error sources that may compromise NAEP findings, including sampling, fair and consistent application of exclusion criteria, and compliance with other aspects of administration procedures. The need for this kind of ongoing audit function is clearly heightened by the expansion of NAEP to provide State-by-State comparisons.

As set forth in Recommendation 12, an ongoing evaluation function should be established, independent of the NAEP contractor, which would regularly examine the overall accuracy of the assessment, assist in

distinguishing real from artifactual patterns and changes in achievement, identify design problems, and if necessary, provide some basis for analytical adjustments to compensate for planned procedural changes as they are implemented, and not retrospectively. This statistical evaluation could also consider issues of subpopulation bias, possibly uneven student motivation, and other factors that might detract from the validity of NAEP findings. Where feasible, NAEP should also be linked to other sources of information on achievement.

Summary of Individually Authored Papers

This report is based on separately authored papers by nearly all panel members. These papers represent the positions of their individual authors, but reflect the deliberations of the entire panel. They provide detail and arguments in support of the panel's findings.

Herbert J. Walberg's paper, National Assessment for Improving Education: Retrospect and Prospect, establishes the policy context for our examination of NAEP, and the importance of its continuation. It places the current assessment in its historical context, and sketches some bold ideas for the future.

Jeanne S. Chall's paper, Could the Decline Be Real? Recent Trends in Reading Instruction and Support in the U.S., places the results of the 1986 reading assessment in the context of long-term patterns and trends, and argues that at least part of the decline may be attributable to changes in methods of reading instruction, especially a too-early emphasis on higher cognitive processes, as well as to less support for reading and remediation in the school, home, and community.

Larry V. Hedges' paper, The NAEP/ETS Report on the 1986 Reading Data Anomaly: A Technical Critique, reviews the technical report on the 1986 reading anomaly by Beaton, et al., and evaluates the evidence presented concerning various hypothesized explanations. He criticizes the strategy of asking whether each hypothesis in turn could explain the bulk of the decline at age 9 or age 17, and suggests that a combination of changes in administration procedures might account for a substantial proportion of the changes in reading performance.

Janet Baldwin's paper, Reading Trend Data from the National Assessment of Educational Progress: An Evaluation, reviews the quality of NAEP reading trend data. She finds that changes in procedures and in test content have confounded the meaning and interpretability of NAEP trend data, especially in the 1984 and 1986 assessments. Dr. Baldwin recommends a more rational framework for identifying NAEP objectives and suggests ways to improve the consistency in score meaning over assessment cycles.

Tej Pandey's paper, Mathematics Trends in NAEP: A Comparison With Other Data Sources, compares NAEP mathematics trends over nearly two decades with findings from the SAT, ACT, ITBS, ITED, GED, NLS-72, and HS&B. He finds no evidence of inconsistencies in the directions of changes between NAEP and other data sources, although the magnitudes of changes are difficult to compare from one test to another where standard deviations are not reported. Dr. Pandey recommends improvements in the taxonomy of content and process categories used in defining NAEP mathematics objectives, and cautions that if NAEP approaches the status of a "national test," then the choice of content for NAEP will influence school curricula.

William H. Schmidt's paper, Quality Control: The Custodian of Continuity in NAEP Trends, addresses the importance of procedural as well as statistical and sampling consistency in NAEP. Dr. Schmidt places the

1986 reading anomaly in the context of other difficulties caused by past procedural modifications, and calls for better quality control mechanisms based on systematic procedures for considering all changes from one exercise cycle to the next.

David E. Wiley's paper, Assessment of National Trends in Achievement: An Examination of Recent Changes in NAEP Estimates, pursues two lines of investigation of the 1986 reading anomaly. First, Dr. Wiley examines both levels and distributions of scores at all three age/grade levels, and finds that, as a consequence of increased variability of the score distributions from 1984 to 1986, at sufficiently low levels of performance, there were declines for all three groups, while at sufficiently high levels, there were improvements for all three groups. Second, Dr. Wiley compares the content of the age 17 NAEP exercises to that of the SAT. While noting important differences, he finds sufficient parallelism to support cautious comparisons of SAT reading performance and NAEP reading performance for high-ability students. He finds that SAT changes do parallel NAEP changes. Dr. Wiley concludes that the magnitude of the NAEP reading scale score changes between 1984 and 1986 together with the large increase in score distribution variability make methodological changes between the two assessments the most likely primary cause of the decline.

Mark D. Musick's paper, Management and Administration of a State-NAEP Program, recommends that the State-NAEP program be established as a program unit within National-NAEP. Dr. Musick considers a range of issues in the administration and governance of such a unit, and in the articulation of State-NAEP with National-NAEP, including instrumentation, sampling, identification of students excluded from testing, local options for expanded assessments within States, test administration, reporting of findings, and other matters. He concludes that establishing and administering a nationwide student testing program that uses the NAEP to provide information on a State-by-State basis is a manageable task.

R. Darrell Bock's paper, Recommendations for a Biennial National Educational Assessment, Reporting by State, provides a comprehensive and detailed technical plan for a National Assessment permitting State-by-State comparisons, and permitting an orderly evolution as statistical methodological advances ("updateability"). Dr. Bock's design for an assessment addresses issues of sampling, assessment cycles, domain definitions, assessment instruments, background questionnaires, administration procedures, scoring, reporting and technical support. His design includes both objective questions and writing exercises, and allows for the linkage of existing State assessments to a national assessment.

John T. Guthrie's and Susan R. Hutchinson's paper, Objectives for State Assessments by NAEP, considers the content of State-NAEP assessments in the light of the purposes these assessments will serve. In specifications for NAEP exercises, the authors argue that ancillary, or unintended, requirements of exercises must be considered, as well as intended objectives. For example, readability of exercises (other than those designed to test reading) should be at least 2 years below the age/grade level at which the exercises are to be used, and inference items

should not depend on factual knowledge that cannot be presumed to be nearly universal among the group tested. The authors also consider whether content and processes should be assessed separately or in concert.

Joan Boycoff Baron's and Pascal D. Forgione, Jr.'s paper, Collecting and Profiling School/Instructional Variables as Part of the State-NAEP Results Reporting: Some Technical and Policy Issues, presents issues and recommendations relating to the collection of NAEP background data. The authors propose criteria for selecting background questions based on prior theory, research, and empirical findings, and propose a long term NAEP research agenda to improve and stabilize the NAEP background data collection.

Leigh Burstein's paper, Reporting State-Level NAEP in a Fair and Credible Manner, highlights the technical and policy issues in reporting State-by-State results that need to be considered in the 1990 and 1992 trails and beyond. Dr. Burstein discusses the purposes and principles that should guide State-level analysis and reporting, and the alternative bases for comparing States that might be examined during the trial cycles. The panel's recommendations are linked with corresponding recommendations of the CCSSO-directed NAEP Planning Project, and with analytical options explored by Haertel for State-level comparisons based on procedures States have used to compare test scores for districts or schools. Burstein concluded that the trial cycles should generate a wide variety of State-level reporting systems. Through ensuing discussion and debate, these would evolve into the core reporting methods for a fully operational State-NAEP.

Edward Haertel's paper, Within-State Comparisons: Suitability of State Models for National Comparisons, first considers problems of equity or fairness that arise with any system for adjusting scores or setting different expectations for different schools, districts, or States. He then describes systems used in several States for reporting school- or district-level achievement, and considers the applicability of these methods for purposes of State-by-State comparisons. In conclusion, Dr. Haertel suggests three possible approaches, including the reporting of unadjusted means for demographic subgroups, comparisons of each State's performance to a predicted level derived from models for subunits (e.g., communities) within the State, and a method using "floating comparison groups," as recommended by the CCSSO Consensus Planning Project.

In summary, both collectively and individually, we have given careful attention to the three issues we were charged to address. It is our hope that our report, recommendations, and conclusions will help to guide and improve the National Assessment of Educational Progress in years to come.

References

Alexander, L., & James, H. T. (1987). The Nation's Report Card: Improving the assessment of student achievement (Report of the Study Group). Washington, DC: National Academy of Education.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1988). Who reads best? Factors related to reading achievement in grades 3, 7, and 11 (Report No. 17-R-01). Princeton, NJ: Educational Testing Service.

Beaton, A., Ferris, J. J., Johnson, E. G., Johnson, J. R., Mislevy, R. J., & Zwick, R. (1988). The NAEP 1985-86 reading anomaly: A technical report. Princeton, NJ: Educational Testing Service.

Chall, J. S. (1983). Literacy: Trends and explanations. Educational Researcher, 12(9), 3-8.

Chall, J. S. (1986a). School and teacher factors and the NAEP reading assessments (Paper commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report, The nation's report card.) (ERIC Document Reproduction Service No. ED 279 667.)

Chall, J. S. (1986b). New reading trends: The NAEP report card. Curriculum Review, 25(4), 42-44.

Koretz, D. (1986). Trends in educational achievement. Washington, DC: Congressional Budget Office.

Koretz, D. (1987). Educational achievement: Explanations and implications of recent trends. Washington, DC: Congressional Budget Office.

Murnane, R. J. (1987, April). Improving education indicators and economic indicators: The same problems? (Paper presented at the meeting of the American Educational Research Association, Washington, DC.)

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC: The Commission.

Part 3. Commissioned Papers

National Assessment for Improving Education: Retrospect and Prospect

Herbert J. Walberg
University of Illinois at Chicago

Congress passed legislation in 1867 to create the U.S. Office of Education which was chiefly to collect statistics with a view toward improving education in the Nation. Today it is increasingly apparent that accurate, comprehensive data are necessary for raising educational productivity and helping to increase the quality of our national life.

The National Assessment of Educational Progress (NAEP) is and should be a major informational vehicle for accomplishing these purposes. Sometimes called "the Nation's Report Card," NAEP was created in 1969 to obtain dependable data on the status and trends of achievement in a uniform, scientific manner. Today NAEP remains the only regularly conducted national survey of achievement at the elementary, middle and high school levels (young adults have also been sampled). Unlike the longitudinal projects High School and Beyond and NELS-88 that follow the same students longitudinally over time, NAEP is a periodic series of cross-sectional surveys of the successive groups of the students of the same age.

The subject areas assessed most often include reading, writing, mathematics, science, and social studies, although citizenship, computers, literature, art, music, and career development are also assessed. As of 1988, NAEP had tested about 1.3 million young Americans--making it one of the largest social surveys ever conducted and the costliest, most comprehensive, and long-standing educational survey in the U.S. and perhaps the world.

NAEP is designed with advice by teachers, subject-matter experts, and citizens with a variety of points of view and representing various constituencies. By a process of consensus, they suggest the design objectives for the subject areas and specify general goals that students should achieve by the three ages usually tested. 9, 13, and 17, and now grades 3, 7, and 11. After review and revision, these specifications are turned over to item writers, who develop questions and other exercises appropriate to the objectives.

The items are reviewed for appropriateness and possible bias, then field tested, revised, and administered to stratified, multi-stage probability samples. The resulting data are analyzed in various ways and then disseminated. The general purpose of NAEP is to provide information that will help educators, legislators, and others to improve education.

This paper discusses NAEP evolution and several future prospects for accomplishing its long-term general purpose. It draws not only on the national discussion of assessment and educational reform but also my

experiences analyzing NAEP data; sitting on consultative panels for both NAEP contractors; advising educational authorities in State and Federal Government, and foreign ministries of education; and serving on technical and policy groups for the International Association for the Evaluation of Educational Achievement and the Organization for Economic and Cooperative Development.

A Retrospective on NAEP

At the 1967 meetings of the American Educational Research Association (AERA), we first heard some of the early arguments about whether or not we should have a national assessment, and, if so, about how it should be designed. Harvard dean of education Theodore Sizer had awakened my interest in the topic; he spoke actively and eloquently for the assessment. Apparently, however, several professional education associations and State authorities resisted the idea because they feared that schools and States would be identified. A concession was made that only broad regions and subgroups of students would be compared, and the emphasis would be on changes over time. The situation, of course, is far different today since a large number of governors, legislators, and interested citizens want to study specific comparisons of States and to evaluate the reforms made in this period of extraordinary ferment.

In the 1970s, NAEP began to extend its scholarly usefulness significantly when the National Science Foundation explored with several scholars the idea of using NAEP science data for "secondary" studies going beyond trend analyses. The University of Illinois at Chicago was awarded a grant with subcontracts to two other universities and to the Education Commission of the States that conducted NAEP at the time. The project converted the massive NAEP data files to a uniform format with control statements for the Statistical Package for the Social Sciences that would make it easier for "secondary" or non-NAEP investigators to analyze the data. The project also produced sample analyses, and trained a group of other secondary users from about twenty universities and research agencies.

This grant resulted in a number of publications and an active special interest group of secondary analysts of NAEP data within the American Educational Research Association. NAEP analysts fall into three classes: subject matter specialists who determine how students perform on individual items and clusters as related to curriculum policy; psychometrists interested in item response patterns; and researchers such as myself who try to determine what home, school, and other conditions students appear optimal for educational achievement so as to suggest possible changes in educational policies and practices. The NAEP contractor, Educational Testing Service, has reported item results and scores by region and types of students, and has recently begun some causal analyses to suggest policies.

In my view, NAEP has extended its utility considerably by enabling secondary users to purchase data tapes at low cost and to analyze the immense bank of data that has accumulated. I have been glad to see this at

first hand both as a long-standing analyst of NAEP data and as an advisor to the former and present contractors, Educational Commission of the States and to Educational Testing Service. At little extra monetary cost and human time, many opinion and "background" items on home, class, and school characteristics and conditions are now given to students, teacners, and principals.

Such items need not be given to entire samples but only random fractions of students taking the achievement tests. They allow us to study how educational conditions and practices, as well as test scores, are changing over time. Analyses of such items are in the tradition of the General Social Survey conducted by the National Opinion Research Center that assesses public opinion. "Core items" are used in every survey to measure trends in opinion or conditions of enduring importance. "Piggy-back" items may be added temporarily for one or more surveys to detect important short-term trends, or those of interest to particular analysts or project sponsors. In these ways, NAEP has become an even greater national asset by providing additional valuable information efficiently since the cost of adding items is small relative to the fixed costs of administration, sampling, data collection, computer processing, and archiving.

Causal Inference

The opinion and background items of NAEP also allow a degree of causal inference. They may, for example, allow us to infer that leisure-time reading and homework enhance achievement, other things being equal, since students who engage in these activities achieve better than students alike in other respects who do not. Of course, causal directionality is uncertain since, for example, motivation may cause leisure reading, homework, and achievement, and achievement itself may cause the other things, even though attempts are made to control all the variables in regression and other analyses.

NAEP, of course, is a periodic survey of cohorts rather than a longitudinal study that follows the same students over time. Longitudinal studies are better designed to detect variations in learning and other personal characteristics during and after schooling attributable to educational practices and conditions. Experimental studies of smaller groups of students randomly assigned to conditions and control groups may even be stronger indicators of these effects.

Neither longitudinal studies nor experiments, however, are infallible: Some members of longitudinal studies refuse to participate or can no longer be located; and they may differ considerably from the original sample. Despite efforts to statistically control the alternative causal variables, analyses of longitudinal data may suggest spurious effects that cannot be completely ruled out--no matter how much sociologists might hope. Experiments that psychologists prefer yield stronger causal confidence since groups differ only in random assignment and the conditions are closely observed, but they may be criticized as studies of how people act under contrived conditions rather than in the real world.

For purposes of causal inference, NAEP can partly compensate for the weaknesses of longitudinal surveys and experiments since it comes closest to estimating policies and conditions in the Nation rather than those encountered by samples remaining from longitudinal surveys and the small, idiosyncratic samples usually obtained in the usual small-scale, single-site experiments. Some effects such as student motivation, the amount of instruction, its quality, homework, and the curriculum of the home seem powerful enough to be detectable in longitudinal, cross-sectional, and experimental studies (Walberg, 1986). So each of these approaches has something to contribute--not only in assessment but in suggesting improvements in practice.

State, International, and Other Comparisons

The original intentions of NAEP were to accumulate data that could be compared with future data to measure progress and to compare broad regions of the country and sub-classes of students such as boys and girls, and members of various ethnic and social groups. These may be worthwhile, but they are far from telling us all we need to know and what might be gleaned from an upgraded NAEP. Of course, we want to know how the Nation is progressing in achievement; but we are even more interested in knowing how it compares with the status and progress of other countries. Policymakers in education want to know why achievement has changed and how to raise it. International cooperation on such matters is mutually beneficial since the generally larger variations among countries than within any particular country make causal effects more detectable.

Even so, the U.S., like Australia, Canada, and West Germany, has no centralized ministry of education; it could be said that we have 50 or more ministries. If State policies led to increased science achievement in Vermont and decreased it in New Hampshire, we could detect neither the differences nor causes if the results were averaged by regions or the whole Nation. Some State governments are paying a greater share of the costs of education than previously, and they have initiated different reforms--with more radical choice plans on line. Some very much want to know how schools compare within their States as well as how their State averages compare with those of other States. The States can learn from one another's experience, and so can the Nation as a whole.

The new demand for results and their measurement, of course, began with A Nation at Risk, the report of the federally-appointed National Commission for Excellence in Education and the several dozen other reform reports that followed it. The National Commission pointed out the poor performance of U.S. students by international standards and the potential contributions of education to national prosperity and welfare. As a result of the reform reports, considerably more money was spent on education and many reforms were made; but it does not seem clear that education is yet as efficient as it should be, and that funds are wisely spent.

If anything, it seems likely that there will be greater demands for information and reforms. At the request of President Ronald Reagan, for example, Secretary of Education William Bennett issued a 5-year followup

on the National Commission report in the spring of 1988. H's evaluation of education focused public attention on what else should be achieved in the Nation's schools. Data released on science achievement in March, 1988 by the International Association for the Evaluation of Educational Achievement (IEA) and the National Science Foundation again showed comparatively poor U.S. achievement and rankings near the bottom of affluent countries and rivaled by several developing countries.

Even more pointedly, the National Governors' Association (NGA, 1986) issued the bluntly titled A Time for Results which calls for higher achievement and deeper and wider reforms. Some recommendations in their report on achievement comparisons, school-site management, parental involvement, governance, diversity, magnet schools, and choice of attendance were not widely considered a decade ago; but are now being enacted in many districts and States. In Results in Education, 1987, the NGA called for indicators that reflect State educational goals; measure higher-level skills demanded by society; and meet the information needs of educators, policymakers, and the public. In addition, the Council of Chief State School Officers voted to compile State indicators including achievement.

At one time, failure rates on the Selective Service Examinations for the military draft were available by State. The State failure rates ranged widely from less than 1 percent in Minnesota to about 35 percent in one Southern State, and they were related to various educational conditions (Walberg, 1979). In recent years, Wall Charts have been issued showing average student performance on the Scholastic Achievement Test and the American College Test and other educational indicators. Since these tests have been taken by selective, non-random, and varying fractions of age groups within States and across time, they are less desirable as indicators than NAEP scores. But widespread publicity and comments about State and international achievement show the enormous public and professional interest in comparisons.

The prospect of better State comparisons, as well as school and international comparisons, comes at the right time. NAEP can help fulfill this important national interest as suggested in the Alexander-James report of the National Academy of Education. But it might also accomplish several less obvious purposes in the long range. Consider several possibilities:

NAEP by Tailored Testing

NAEP might change this century's convention of giving each child within a class or grade the same test which is similar to old-fashioned "batch processing" in industry. A far more efficient and time-saving approach is "tailored-testing" (see Carroll, 1982) which flexibly adapts test items to students over great ranges of ability.

For several decades, it has been possible both in principle and in practice to program computers to assign the most discriminating items to each student, based upon her or his prior responses during the testing session. In fact, the idea goes back to the origins of mental testing:

Alfred Binet, of course, administered intelligence test items of a given difficulty to children depending on how well they did on the first few items tried (Carroll, 1982). As few as ten such tailored items can yield scores as reliable as many more batched items suited to the average student. Alternatively, an hour or two of tailored items might yield accurate individual assessments not in one subject but in all the major subjects of the standard curriculum. Or, in the same time, such items could provide highly detailed assessments of skills in a single discipline, for example, word choice, grammar, spelling, and punctuation in written composition. From such results, tailored instruction could avoid teaching what students already know and what they are yet incapable of learning until they meet prerequisite skills.

The increased efficiency in time use and the computer's capacity to record large amounts of information make it feasible to monitor individual student progress more frequently, accurately, and comprehensively. With a thorough, continuing assessment of what each student needs to learn, it should become equally feasible to provide computer-adapted or tailored instruction. Such instruction is by no means a panacea, but it is among those educational methods that provide moderately superior achievement; and it has the further advantage of saving students' study time (Walberg, 1986). It can be expected that computer costs will continue to fall, while software increases in sophistication and interest.

NAEP by Modem

NAEP itself could in a decade or two be done by computer hookups and this possibility seems worth exploration. In principle, it would be feasible to conduct sample surveys of districts, schools, and students directly by computers. Students, for example, could rapidly complete tailored tests and questionnaires by terminal and modem. In compensation, students and schools could receive an instant report on the results. Now, of course, they receive nothing, except perhaps a newspaper report a year later if they chance to come across it. The further advantage of a national hook up is the speed at which surveys and tests can be completed. The time-consuming steps of printing tests and questionnaires, mailing, scanning and screening data, and the like could be skipped. Even analyses could be automated, and produced at electronic speed.

Like national polls of 1,500 respondents that provide reasonably accurate estimates of public opinion in the Nation, direct sampling by telephone controlled by computer might make NAEP fast and cheap. Quarterly or even monthly survey reports on important output measures could be made routine as they are in commerce and industry. Local, State or national assessments of special topics might be commissioned and completed in less than a month. In principle, we would not have to wait a year for the Kappan's Gallup poll on education, several years for cycles of the present National Assessment nor as much as a decade between international comparisons. We cannot follow the rapid reforms in the States if the information is obsolete before it is processed.

Federal Statistics

Federal Government spending on education statistics is small by several standards. In school year 1982-83, for example, spending on public elementary and secondary schools in the U.S. by Federal, State, and local government was respectively $56, 52, and 8 billion, which totals $116 billion, or 4.5 percent of the $2.6 trillion national income (Indicators, 1985, p. 22). If the Federal Government spent $100 million on better educational statistics, it would amount to less than one-tenth of one percent of total educational spending on public schools and might increase efficiency by many billions.

Given U.S. Government spending of $1.4 billion on statistics (Alonso and Starr, 1985, 123), education's 4.5 percent share (based on the public school percent of national income) would be $63 million, in contrast to $8.7 million in current spending by the National Center for Education Statistics. Higher spending on research should yield better statistics and make the "education industry" more comparable to agriculture, medicine, and various industries that base practice upon productivity comparisons.

Conclusion

The present seems a time for great opportunity in educational reform and research in education. Agriculture, engineering, and medicine made great strides in improving human welfare as doubts arose about traditional, natural, and mystical practices, as the widened measurement of results intensified, as experimental findings were synthesized, and as their theoretical and practical implications were coordinated and vigorously implemented and evaluated.

Education is no less open to humanistic and scientific inquiry and no lower in priority since half the workers in modern nations are in knowledge industries, and the value of investments in people is now more apparent than ever (Walberg, 1983). Although it is possible to find fault with federal statistics on education, the last decade or two has been a period of quiet but significant accomplishments; and larger amounts of valuable data are being accumulated, and put to good use by policymakers. More is to come.

NAEP has become a national asset. It can serve as a sturdy benchmark for our accomplishments and failures at reform; and policymakers need to know about both. The technical problems raised by expansion of NAEP to provide State estimates are well within current technology.

A State-level NAEP can be accomplished at reasonable cost, using existing technology, in a way that assures the preservation of national trend estimates. NAEP can also provide "meter sticks" or links for the States to make district and school comparisons within States. NAEP can be also made more useful by other means in the longer range by modernizing its technology, although extreme caution is required since NAEP's main value is to provide benchmarks and indicate trends.

The reading anomaly hardly implies that national and State assessments are unmanageable. It appears that the present NAEP contractor may have made too many well-intentioned procedural changes too quickly; but, as pointed out by Jeanne Chall and others, actual declines cannot be ruled out. Studies now underway should help to resolve the question of how much of the observed reading decline from 1984 to 1986 reflected procedural and actual changes. Anomalies and other uncertainties can be minimized or avoided, as suggested by the committee, by carrying out new procedures only for very strong reasons, and exploring them sparingly and simultaneously with the old procedures to test their efficacy before substantially committing NAEP and risking its main values.

In my opinion, nonetheless, neither NAEP nor the testing profession in general can afford to slacken efforts to innovate and implement superior technologies. Indeed, as discussed above, there may be as much need for testing reform as changes in other educational practices. Calibrated "meter sticks" for test equating, for example, would facilitate comparisons across time and place; these would simplify scientific studies and also provide more comparable and comprehensible information to the public. Tailored testing can save immense amounts of student time and yield more accurate results. In principle and to some extent in practice, computer-assisted testing including telephone hookups can cut time for testing feedback to students, teachers, and policymakers by as much as 90 percent or more; and the value of such information is proportional to its speed. The testing profession, moreover, can contribute much more to the Nation's need for increased achievement by coordinating tests more directly with curriculum and instruction.

NAEP should not be jeopardized by unproven techniques, as explained above. But neither should we divorce it from technical progress. NAEP, for example, has already greatly added to its utility by cautiously and unobtrusively adding "background" items on educational practices and conditions. This addition has allowed many useful policy analyses by NAEP staff and outside researchers.

Although NAEP's main effort and most of its resources should go into proven methods that worked well in the past, NAEP might selectively lead or be used in the development of innovations and new technologies. These efforts need not be carried out by NAEP itself. Rather State and local educational specialists, as well as independent investigators are carrying out many of the newer testing practices even now, and they can make good use of the valuable collection of NAEP items and statistical information on them. In these ways and by increased cooperation, the testing profession through NAEP and other projects can contribute much to meeting the challenge of increasing what our students learn.

References

Alonso, William, and Starr, Paul (1985). A nation of number watchers. Wilson Quarterly, 9 (3), 93-123.

Carroll, J. B. The measurement of intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence. New York: Cambridge University Press, 1982.

Cooke, Charles, Ginsburg, Allen, and Smith, Marshall (1985). The sorry state of education statistics. Madison: Wisconsin Center for Educational Research.

Feistritzer, C. Emily. (1985). Cheating our children. Washington, DC: National Center for Educational Information.

Harnqvist, Kjell (1984). An empirical study of long-term effects of education. Tel Aviv, Israel: Paper presented at the First International Conference on Education in the 1990s.

Indicators of Education Status and Trends (1985). Washington, DC: U.S. Department of Education. Leontief, Wassily, Duchin, Faye, and Szyld, Daniel B. (1985) New Approaches to Economic Analysis. Science, 112, 419-427.

National Governors' Association (1986). A time for results. Washington, DC: NGA.

National Governors' Association (1987). Results in education, 1987. Washington, DC.: NGA.

Raizen, Senta A. and Jones, Lyle V. (1985). Indicators of precollege education in science and mathematics: A preliminary review. Washington, DC: National Academy Press.

U. S. Commission on Excellence in Education (1983). A nation at risk. Washington, DC: U. S. Government Printing Office.

Walberg, Herbert. J. Scientific literacy and economic productivity in international perspective. Daedalus, 1983, 112, 1-28.

Walberg, Herbert J., and Rasher, Sue Pinzur (1979). Achievement in the 50 states. In Herbert J. Walberg (Ed.), Educational environments and effects. Berkeley, CA: McCutchan Publishing.

Walberg, H. J. (1986). Synthesis of research on teaching. In M. C. Wittrock (Ed.), Handbook of research on teaching. New York, NY: Macmillan.

Could the Decline Be Real?
Recent Trends in Reading Instruction and Support in the U.S.*

Jeanne S. Chall
Harvard University

The purpose of this paper is to present evidence on the probability that the declines in the NAEP 1986 reading scores are real and not primarily an anomaly. I will do so through an analysis of the trends in NAEP reading test scores from 1971 to 1986. I will further relate the NAEP reading scores to the reading instructional practices and to the resources available to each cohort prior to the time it was tested.

Before analyzing the reading score trends from NAEP, it is well to present briefly what is known about the influence of the school, home and community on reading achievement.

Influence on Reading Achievement

1. Reading instruction, including textbooks, and reading resources in the school, home and community have significant effects on reading achievement. (See, for example, Chall, 1986a and 1987; Chall & Snow, 1982 and 1988; Coleman, Campbell, Ho[illegible]rtland, Mood, Weinfeld, & York, 1966; R.L. Thorndike, 1973.)

Some school factors have significant effects on students of all ages and grades while many seem to be important at early or later stages of reading development. For example, the research over the past 70 years has found that direct instruction and practice of word recognition and phonics in grades 1 and 2 (and later for older students still reading at these early levels) produces better achievement in word recognition and comprehension. This advantage is cumulative and tends to be found in the scores in later grades (Chall, 1967 and 1983a; Williams, 1986; Perfetti, 1985; Anderson, Hiebert, Scott, & Wilkinson, 1985, What Works, 1986; Bennett, 1986). Similarly, the direct teaching of reading comprehension strategies and word meanings has significant effects at the intermediate grades and higher, when the basic word recognition and decoding skills are mastered, and when the reading materials go beyond the familiar and known, contain specialized and abstract words, and the texts require critical comprehension strategies (Gray & Holmes, 1938; E.L. Thorndike, 1917; Chall & Snow, 1982 and 1988).

2. Reading changes in characteristic ways as it develops--from its beginnings (and prebeginnings) to more advanced levels (Chall, 1979 and

* I wish to thank Edward Haertel for his helpful critical comments and detailed informative charts on the probable trends in the 1971 to 1986 NAEP reading scores, and Mary Curtis and Sue Conrad for their reactions to an earlier draft.

1983b).

If reading is divided into levels or stages, a major break seems to come at about grade 4. Pre-grade 4 reading rarely goes beyond the language and knowledge that the reader has through listening, direct experience and TV. Reading beyond grade 4 generally deals with texts that go beyond what is already known--texts that are ever more complicated, literary, abstract and technical. And these texts require of the reader more world knowledge, ever more sophisticated language and cognitive abilities to engage in the interpretations and critical reactions required. Materials of 4th grade level and beyond are more difficult in content, in linguistic complexities, and in cognitive demands.

Reading at the earlier levels (approximately 1st through 3rd grade reading levels) requires proficiency in word recognition, decoding and fluency. If the reader's native language is English, it requires relatively little "stretching" of linguistic and cognitive abilities. Beginning with the intermediate reading levels (approximately 4th grade reading level and beyond), the challenge becomes primarily linguistic and cognitive. However, without the fluent recognition of words, linguistic and cognitive skills cannot function in reading comprehension (Perfetti, 1985; Anderson, Hiebert, Scott, & Wilkinson, 1985; Williams, 1986; LaBerge & Samuels, 1976).

3. These stages or levels are generally cumulative and continuous. If word recognition lags behind, comprehension will lag behind, even though the meanings of the words are known and the ideas are understood when heard. Accurate and automatic recognition of words is necessary for efficient reading and comprehension. Thus, instruction that focuses on the developmental changes in the reading process will be more effective, other things remaining equal. (LaBerge & Samuels, 1976; Chall, 1983b; Perfetti, 1985).

4. Compensatory education and remedial services for children at risk and for those who are having difficulty will improve reading achievement (L.C. Smith, 1979; Kraus, 1973). The earlier children are given the remedial help that they need, the more effective it is.

Among other school and teacher factors having significant effects are: teacher excellence, time on task, optimal difficulty level of materials of instruction, frequent assessment to guide instruction (Chall, 1986a).

5. Home and community factors have long been recognized as contributing to reading achievement. In recent times, the Coleman report (1966) and the IEA International Study of Reading Comprehension (R.L. Thorndike, 1973) have focused on family background as the major factor in reading achievement.

6. Among home and community factors related to reading achievement are: reading materials in the home (the more, the better); television (the less, the better); education of parents (the more, the better); and homework (the more, the better) (see NAEP, 1985).

The Relation Between NAEP Scores and Changes in School and Home Influences

In the remaining part of this paper, I will attempt to show how changes in school and home factors, such as those noted above, may have contributed to the increases and declines in the NAEP reading scores. Although my main concern is with the 1986 reading scores, I will also analyze the NAEP reading scores for 1971, 1980, and 1984. I do so because my earlier analyses of NAEP reading scores and the school and home factors to which students were exposed seemed to explain the increases and decreases among the different age groups, as well as the gains and losses over different time periods (Chall, 1983c, 1986a, and 1986b).

I begin, first, with the school and home influence on reading for the different time periods.

It should be realized that only the broadest factors in school, home and community influences could be considered for this essentially macroscopic analysis. Further, trends in the conceptualization of reading instruction and support for beginners receive major attention, since beginning reading affects not only achievement in the early grades but also in later grades (Chall, 1967 and 1983a).

Prevailing Views and Practices in Reading Instruction: 1920 to the late 1960s

From about 1920 to about the late 1960s, the major focus in the teaching of reading, from the first grade on, was on "reading for meaning," i.e., on reading comprehension (Chall, 1967 and 1983a; Anderson, Hiebert, Scott, & Wilkinson, 1985). The reading textbooks contained limited vocabularies and relatively little systematic instruction was given in phonics. (See Chall, 1967 and 1983a, for instruction during these decades based on basal readers and their accompanying teacher's manuals.)

Views on Reading in the 1970s (from the late 1960s to the late 1970s)

During the 1970s (as well as the late 1960s), there was a change in most beginning reading programs. Although they still had children recognize whole words (sight) and read for meaning in grade 1, there was an earlier and more systematic emphasis on phonics or decoding (Popp, 1975). As a result, the basal reader textbooks became harder, i.e., they contained more different words, grade for grade. The methods textbooks for teachers published in the 1970s also paid more attention to teaching phonics and decoding (Chall, 1983a). Formal reading instruction was also begun earlier than in the 1960s. Many schools started to teach reading in kindergarten.

More resources seemed to be available for those at risk--for disadvantaged urban preschoolers in Head Start, and school age children in Title I (later Chapter 1). Children of all social levels with learning disabilities received remedial help under Federal Law 94-142. Sesame Street and The Electric Company provided informal stimulation for learning to read in the home.

The emphasis on getting the beginner off to a good start was implicit also in the "basic skills" movement of the 1970s, which was concerned particularly with urban children who were not achieving academically. Somewhat later came the effective schools programs, which also focused on beginning reading. The overall reading emphasis of the 1970s could be characterized as giving children an early and strong beginning.

NAEP's recent report confirms our characterization of reading in the 1970s: "The decade of the 1970s in particular was an era of emphasis on the 'basics'..." (Applebee, Langer, & Mullis, 1987, p. 35).

Reading Views and Practices in the 1980s

The late 1970s and 1980s saw another shift in reading. Reading research, theories, and practices began to focus on reading comprehension, particularly higher-level comprehension processes. A parallel shift took place in the teaching of writing, with an emphasis on the writing process and higher cognitive processes while writing (Applebee, Langer, & Mullis, 1987).

This broad characterization of reading in the 1980s is confirmed by Applebee, Langer, and Mullis (1988) who wrote: "Recommendations for good teaching include ... greater emphasis on comprehension strategies" (p. 5).

It is hypothesized that these trends in reading instruction, particularly in the reading instruction and resources in the early grades, are related, at least in part, to the NAEP scores, especially to the scores for the 9-year-olds.

Trends in NAEP Reading Scores

NAEP Reading Scores for 1971 to 1980

Table I, adapted from NAEP, presents the reading gains from 1971 to 1980 for three age groups--on literal comprehension, inferential comprehension, and reference skills. It contains, in addition, the year in which the different cohorts were in grade 1.

From the table, we note that younger children, the 9-year-olds (4th graders), made the greater gains, more than the 13- and 17-year-olds.

Why the differences? A likely hypothesis lies in the changes in reading instruction that began in the late 1960s and continued through the 1970s--changes characterized above and directed more to younger children than to older students. Thus, the 1980 4th graders benefited from an earlier start, from more and earlier phonics, harder basal readers grade for grade, more home instruction and stimulation through Sesame Street and The Electric Company, more remedial help to those who needed it, Head Start, and Chapter 1. The significant gains of the 9-year-olds on all reading subtests in 1980, as compared to 1971, may be attributed, at least in part, to their stronger reading instruction--reading instruction that "matched their needs"--as well as to their stronger reading environment.

Table I

National Assessment of Educational Progress
Gains in Reading Scores, 1971 - 1980

	Grade 4 (Age 9)	Grade 8 (Age 13)	Grade 12 (Age 17)
Total Reading	+3.9*	+0.8	-0.9
Literal Comprehension	+3.9*	+1.6*	-0.2
Inferential Comprehension	+3.5*	-0.6	-2.1*
Reference Skills	+4.8*	+0.9	+0.8

Grade 4 Scores	Grade 8 Scores	Grade 12 Scores
1971 cohort, in Grade 1, 1967	1971 cohort, in Grade 1, 1963	1971 cohort, in Grade 1, 1958
1980 cohort, in Grade 1, 1978	1980 cohort, in Grade 1, 1972	1980 cohort, in Grade 1, 1968

Note. From NAEP, 1981. *Significant.

From: Chall, J.S. (1983). Literacy: Trends and explanations. Educational Researcher, 12:9, 3-8.

On the other hand, the 4th graders of 1971 were in 1st grade in 1967, before Sesame Street and The Electric Company and before the schools and textbooks changed toward earlier and stronger beginning reading programs (Chall, 1983a).

The grade 8 (age 13) gains in 1980 suggest significant influences from the cumulative effects of the stronger early reading programs and resources of the 1970s on literal comprehension (see Table I). The lesser effects on inferential comprehension are not surprising since it is influenced more by cognition than by reading skills. Since the 1980 8th grade cohort was in grade 1 about 1972, when some would have been exposed to stronger beginning reading programs and richer home reading environments, the improved scores in literal comprehension may reflect this stronger beginning. The 1971 8th grade cohort, on the other hand, was in grade 1 about 1963, when the beginning reading programs were not as strong.

While the influence of beginning reading programs on 12th graders' reading achievement would be weaker than for 4th and 8th graders, it is significant that both the 1971 and 1980 cohorts were in grade 1 before the 1970s, a time of weaker beginning reading programs.

NAEP Reading Trends: 1980 to 1984

Table II presents NAEP scores for 1980 and 1984 (and also for 1986) for 9-, 13- and 17-year-olds, and Table III presents the gains and losses for

each age. These are the scaled scores from the respective assessments, taken from Table 2-1 in Beaton's report (1987) by Edward Haertel. According to Haertel, the numbers in The Reading Report Card are slightly lower, but the patterns are the same.

According to the scaled scores in Tables II and III, the scores for the 9-year-olds in 1984 leveled off, or tapered; the scores for the 13-year-olds declined somewhat, while the 17-year-olds seemed to have gained somewhat.

TABLE II

Trends in NAEP Reading Scores
1980, 1984, 1986

	1980		1984		1986	
Age	Scores	Year in Grade 1	Scores	Year in Grade 1	Scores	Year in Grade 1
9	215	(1975)	213	(1979)	207	(1981)
13	259	(1971)	258	(1975)	260	(1977)
17	286	(1967)	289	(1971)	277	(1973)

TABLE III

Gains and Losses at Same Ages
from 1980 to 1984; 1984 to 1986

Age	1980 to 1984	1984 to 1986
9	-2	-6
13	-1	+2
17	+3	-12

According to The Reading Report Card (NAEP, 1985), both the 13- and 17-year-olds gained, with a tapering or leveling of the reading scores for the 9-year-olds. In a cohort analysis based on date of birth, The Reading Report Card concluded that the stronger scores of the 13- and 17-year-olds in 1984 could be attributed to their stronger scores when they were 9 years old.

Thus, the 1984 NAEP data confirm the importance of a strong beginning in reading--not only for the 9-year-olds' scores, but for their scores when they reach ages 13 and 17.

The NAEP scores from 1971, 1980 and 1984 tend also to give support to a developmental, multiple-stage theory of reading, rather than a single-stage theory that was prevalent during the decades before the late 1960s and again since the middle 1970s. NAEP's data for 1980 and 1984 suggest the greater effectiveness of a stronger, basic skills emphasis for the younger children, which tends to show up at age 9 and later at ages 13 and 17.

It is also possible that the growing concern for teaching the higher-level cognitive skills beginning at grade 1, which began in the late 1970s and early 1980s, is responsible, at least in part, for the leveling of the 1984 NAEP scores among the 9-year-olds. This is suggested by a general movement toward a greater emphasis on reading comprehension even in grades 1 and 2, following the research emphasis on comprehension which started at about 1975 at the University of Illinois Center for the Study of Reading and at many other universities. Although the comprehension research was generally conducted on students in 4th grade and above, the results were applied to grades as early as kindergarten and grade 1. The comprehension emphasis of the 1980s came also from the whole language movement, which focused on the linguistic and cognitive aspects of reading, right from the start, with little or no concern for accuracy of word recognition and phonics (F. Smith, 1979).

Another hypothesis for the tapering of the age 9 scores is the decline in remedial services in schools nationally. With the various State laws cutting spending (e.g., Proposition 2-1/2 in Massachusetts), the funds for special services in schools were cut. This included remedial reading services, with the exception perhaps of federal funds for children with learning disabilities under PL 94-142. Thus only those children with severe reading problems received the help they needed. Those with milder problems were not given the help and were thus negatively affected.

In summary, the tapering of the 1984 scores for the 9-year-olds seems to reflect a return to an emphasis on reading comprehension for beginners as well as for more mature readers, with a decline in teaching word recognition and decoding and a lessening of resources for reading. The gains of the 13- and 17-year-olds could be attributed to their initiation into reading during the 1970s when beginning reading instruction was stronger, with a greater emphasis on word recognition and decoding.

NAEP Reading Trends: 1984 to 1986

The 1986 scores, as compared to the 1984, present a somewhat different picture. The declines for ages 9 and 17 are greater (about one-half year for 9-year-olds and one year for the 17-year-olds) than during any of the earlier testing periods, which covered longer durations. Further, no appreciable change was found in the scores for the 13-year-olds. (See Tables II and III.)

The position taken by most members of the Technical Review Panel is that the magnitude of the 1986 decline over the 1984 scores is due mainly to anomalies inadvertently introduced in the testing procedures. While it is possible that methodological and procedural anomalies may account for the greater magnitude of the 1986 declines, I propose that the trends in the declines suggest that they may be real, particularly the decline in the scores of the 9-year-olds. Indeed, the 1986 decline among the 9-year-olds seems to continue the leveling trend begun in 1984. Similar to the 9-year-olds of 1984, the 1986 9-year-olds learned to read when the beginning reading programs placed less emphasis on word recognition and decoding and a greater emphasis on reading comprehension and higher cognitive processes. Meyer, Hastings, and Linn (1988), for example, found that the 1986 1st grade basal reading program of a major publisher provided less instruction in phonics than in its 1978 edition. Similarly, Neill (1987) found that the teacher's manuals of the 198[illegible] and 1985 1st grade basal readers devoted less space to teaching decoding and more to teaching comprehension and multiple meaning of words than those published in the 1970s.

Further support for the "realness" of the 1986 decline is gained by examining the score distributions for 1986 as compared to 1984. Beaton (1987) notes that the 1986 distribution at age 9 shows a shift in scores, with a larger proportion of students scoring at a very low level. This would suggest that the instructional shift in the 1980s to a stronger emphasis on comprehension for beginning readers was even less effective for the lower than the average and higher achievers--thus confirming the research of the past 70 years (Chall, 1967 and 1983a).

That the 13-year-olds' scores in 1986 did not decline from those in 1984 is also consistent with the earlier NAEP trends--gains made by 9-year-olds tend to strengthen the scores of the same cohort at age 13. Since the 13-year-olds were in the 1st and 2nd grades in the late 1970s, which were characterized by a stronger emphasis on word recognition and phonics, they were more prepared to benefit from the emphasis on reading comprehension that they may have received when they were in the intermediate and upper elementary grades. While the stable scores for the 13-year-olds tend to support the "realness" hypothesis, they do not seem to support the anomaly hypothesis since no evidence has come forth that the changes in procedures and methods of testing were different for the 13- as compared to the 9- and 17-year-olds.

The 1-year drop, the highest in magnitude, for the 17-year-olds seems to be the only one that does not fit the "realness" hypothesis. An explanation based mainly on the beneficial effects of stronger beginnings does not seem to hold. Since the 1986 17-year-olds cohort was in the primary grades during the 1970s, as was the 1984 cohort, the expectation would be for a gain as in 1984.

Were there other factors that might have contributed to the loss, besides the changes in procedures? I propose, as a possibility, the publication and wide influence of A Nation at Risk (1983) and the other "reform reports," published around 1983 and 1984. Essentially, the reports

concluded that high school students' achievement was inadequate for an information/technology age and that steps must be taken to improve achievement through higher standards and curriculum requirements, more difficult textbooks, and a greater emphasis on the higher mental processes. The reports generally gave little attention to the achievement of students who were already having difficulty meeting the lower standards. Some of the reports suggested remedial instruction for the lowest achievers, but it is questionable if much was provided. (See Chall & Davidson, 1984.)

It is reasonable to expect that the higher standards and increasing difficulty of the curriculum would be of benefit more to students in the upper rather than the lower half in achievement. Since the lower half was already struggling to meet the lower standards, raising standards might have made it even more difficult for them to achieve, unless additional help was provided. Thus, the calls for school reform that were meant to help all to achieve better might have ironically contributed to declines in the achievement of the lower achieving students.

That this might have occurred is suggested by the changes in the frequency distributions of the 1986 NAEP scores. The curve for 1986 17-year-olds is flatter than for 1984. While the 1986 distribution had more high scorers than the 1984, it had an even greater number of low scores (Beaton, 1987). Thus, the 1986 score decline among 17-year-olds may be a phenomenon that occurred mainly among the lowest achievers. This is reasonable when placed in the educational context of the rising standards and expectations for high school students from 1983 on, with instruction geared less and less to the needs of the low achievers. The reform reports were not alone in their focus on the "higher" cognitive processes for all students. During the 1980s reading conferences and reading journals also focused on reading comprehension and the higher thought processes in reading. There was a decline in conference presentations and journal articles on early reading, as if this problem had been solved (Neil, 1988). NAEP, from its first reading assessment in 1971 to the present, has placed first emphasis on comprehension strategies. Further, NAEP's testing begins with age 9 and includes only various types of reading comprehension. Thus, it is difficult even at this early age to sort out the "basic skills" from the higher cognitive processes. The growing consensus among reading specialists that more of the higher reading comprehension processes needs to be assessed and taught may have added further to the mismatch of the reading instruction for the lower half of high school students, who may still lack basic skills.

Hence the hypothesis for the 1986 score decline of the 17-year-olds is similar to the hypothesis for the 1980 decline among 9-year-olds--a mismatch in instruction in light of student needs. The higher curriculum demands and expectations, without the needed remedial supports, were probably less effective for the younger students and for the lower-achieving older students.

Influences from out-of-school factors seem also to have been building up to produce lower scores in 1986. There was a decline in the support of

libraries (Center for Education Statistics, 1987), and an increase in the time spent viewing TV, particularly among the poorer readers (NAEP, 1985).

There was also a question of the positive effects of "higher-level thinking" or "process" for writing as well as for reading. Applebee, Langer, and Mullis (1987), in their analysis of the NAEP 1984 scores, wrote: "...more attention to the process of reading and writing [was] less clearly related to achievement .....Students who reported their teachers emphasized process-oriented approaches...wrote no better than those who reported little or no process instruction" (pp. 35-37). Similar trends were found for reading instruction where "increased use of such teaching approaches as having students answer their own questions about what they read, take notes, and learn how to find the main idea of a paragraph were inconsistently and sometimes negatively related to reading proficiency" (p. 37).

The NAEP report on the 1986 scores, Who Reads Best?, gives further evidence that the higher cognitive emphases, while perhaps more effective for higher achieving students, may not be effective for younger and for older, lower achieving students. Applebee, Langer, and Mullis (1988) note that, when asked what "strategies students might adopt when they found that something was difficult to read...," there was "a shift in strategies between the lower and upper grades."

> Among third grade students, for whom reading is a newer skill, the preferred strategy was to sound out the difficult parts (33 percent), follow it closely by asking for help (22 percent). By grade 11, students were more likely to rely on the meaning of the passage as a whole to help them through the hard parts. (p. 30)

When the upper and lower quartiles were compared for each grade, fewer differences were found at grade 3--"in both groups, sounding out words was the most popular strategy" (Applebee, Langer, & Mullis, 1988, p. 31). The strategies of the poorer readers were somewhat similar across the grades--with a slight increase in the proportion relying on rereading and context and a slight decrease in sounding out words.

These observations, and many others throughout Who Reads Best?, tend to give confirmation to a developmental theory of reading that emphasizes different aspects of reading at different stages of reading development, rather than to a single-stage theory that focuses, from the very start, on viewing and teaching reading as a higher-level cognitive process.

Summary and Conclusions

I have attempted to present evidence on the "reality" of the NAEP 1986 reading score declines. While accepting the possibility that the magnitude of the declines stems from the unanticipated consequences of the changes in procedures and methods followed in administering the 1986 tests, I have proposed that the reality of the declines cannot be overlooked, especially when one relates them to changes in reading instruction and to available resources. When the NAEP reading scores for 1971, 1980, 1984 and 1986 are

viewed in relation to each other and in relation to the reading programs the various age groups were exposed to, one finds higher scores following strong beginning reading programs and strong reading supports. Losses are found when beginning reading is taught as a high-level cognitive process from the start, to 1st and 2nd grade children and to older students still on these beginning levels. There is also evidence that the better results from stronger reading programs in the primary grades contribute to higher scores among these cohorts when they are 13- and 17-year-olds. However, later conditions in the school and community may have stronger effects.

Generally, the probability that the 1986 declines are real is supported by the historical analysis of the trends in scores and the trends in instruction and resources. If the declines reflect the less effective school practices, then "attention must be paid."

The persistence of the low levels of reading proficiency among the 17-year-olds in 1971, 1980, 1984 and 1986 is of great concern to all. According to the NAEP proficiency scale of 1984, only 39 percent of 17-year-olds could read at an "adept level," a level permitting the reading of high school textbooks. Further, less than 20 percent of the urban disadvantaged and 16 percent of black students could do so. This mismatch of reading abilities of the vast majority of high school students with the difficulty of their required texts in school and with general adult magazines like Time and Newsweek and a newspaper like the New York Times helps explain why so many find high school too difficult, irrelevant, or "boring," and choose to drop out. Indeed, the sharpest decline in reading achievement is among 17-year-olds, who are dropping out of school in increasing numbers. Since the lowest achieving 17-year-olds have probably left school before taking the NAEP in the 12th grade, the NAEP may already be an overestimate of the reading of 17-year-olds in the U.S. The decline in the NAEP scores for the 9-year-olds in 1986 is even more serious, however, for it foretells lower scores at ages 13 and 17.

Another cause for concern with the 1986 NAEP reading scores is the change in the frequency distributions for all three age groups, particularly for ages 9 and 17. According to Beaton (1987, p. 36), the 1986 assessment of reading produced more than twice as many 'low scorers' [students below a Basic Level of proficiency, about a 3rd grade level] as the 1984 assessment.

The declines in scores among the lower achievers, in 1986 as compared to 1984, adds further confirmation to the probability that the reading scores reflect, in part, the real achievement of the students, and that these low scores are related to the less effective instructional emphases and the decline in resources.

Whether the above essentially macroscopic analysis will be confirmed by more detailed, microscopic analyses is yet to be determined. My hope is that it will be taken seriously enough to be studied further. We are beginning such a study, which includes estimating the changes in children's reading textbooks, coordinated teacher's manuals, methods textbooks and materials in classrooms.

Although the probability of the "realness" of the 1986 NAEP scores may strike some as pessimistic, it can, I believe, lead to constructive outcomes. For essentially, the declines in the 1986 scores as well as the rise and fall of NAEP reading scores from 1971 to 1984 seem to show that what teachers teach and textbooks "cover," what families provide and communities enhance, do make a difference--and these difference are reflected in scores on the NAEP, and they can be changed for the better.

References

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). Becoming a nation of readers: The report of the Commission on Reading. Champaign, IL: The National Academy of Education and the Center for The Study of Reading.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1987). Learning to be literate in America. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1988). Who reads best? Factors related to reading achievement in grades 3, 7 and 11. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Beaton, A. E. (1987). The NAEP 1985-86 reading anomaly: A technical report. Princeton, NJ: Educational Testing Service.

Bennett, W. J. (1986). First lessons: A report on elementary education in America. Washington, DC: U.S. Government Printing Office.

Center for Education Statistics, (1987). The condition of education: A statistical report. Washington, DC: U.S. Government Printing Office.

Chall, J. S. (1967). Learning to read: The great debate. New York: McGraw-Hill.

Chall, J. S. (1979). The great debate: Ten years later, with a modest proposal for reading stages. In L.B. Resnick and P.A. Weaver (Eds.), Theory and practice of early learning (Vol I). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chall, J. S. (1983a). Learning to read: The great debate (Updated edition). New York: McGraw-Hill.

Chall, J. S. (1983b). Stages of reading development. New York: McGraw-Hill.

Chall, J. S. (1983c). Literacy: Trends and explanations. Educational Researcher, 12:9, 3-8.

Chall, J. S. (1986a). School and teacher factors and the NAEP reading assessments. Paper commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report, The Nation's report card, August 1986. ERIC Document Reproduction Service No. ED 279 667.

Chall, J. S. (1986b). New reading trends: The NAEP report card. Curriculum Review, 25:4, 42-44.

Chall, J. S. (1987). The importance of instruction in reading methods for all teachers. In Intimacy with language: A forgotten basic in teacher education. Baltimore, MD: The Orton Dyslexia Society.

Chall, J. S., & Davidson, R. G. (1987). The quest for higher literacy: Views from six national reports. Unpublished paper, Harvard Graduate School of Education, Cambridge, MA.

Chall, J. S., & Snow, C. (1982). Families and literacy: The contribution of out-of-school experiences to children's acquisition of literacy. A final report to the National Institute of Education, Washington, DC. ERIC Document Reproduction Service No. ED 234 345.

Chall, J. S., & Snow, C. E. (1988). School influences on the reading development of low-income children. The Harvard Education Letter, 4:1, 1-4.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of educational opportunity. Washington, DC: U.S. Government Printing Office.

Gray, W. S., & Holmes, E. (1938). The development of meaning vocabularies in reading: An experimental study (Publications of Laboratory Schools of the University of Chicago, No. 6). Chicago: University of Chicago Press.

Kraus, P. E. (1973). Yesterday's children: A longitudinal study of children from kindergarten into the adult years. New York: John Wiley and Sons.

LaBerge, D., & Samuels, S. J. (1976). Toward a theory of automatic information processing in reading. In H. Singer & R. Ruddell (Eds.). Theoretical models and processes of reading (2nd ed.). Newark, DE: International Reading Association.

Meyer, L. A., Hastings, C. N., Linn, R. L. (1988) Assessing early reading with new decoding and comprehension measures. Champaign, IL: The Center for the Study of Reading.

National Assessment of Educational Progress (NAEP). (1985) The Reading report card: Progress toward excellence in our schools. Princeton, NJ: Educational Testing Service.

National Commission on Excellence in Education (1983). A nation at risk. Washington, DC: U.S. Government Printing Office.

Neil, N. (1987). Analysis of basal readers for district textbook adoptions. (Unpublished paper, Harvard Graduate School of Education, Cambridge, MA.

Neil, N. (1988). Beginning reading, International Reading Association and NAEP: Are there connections? Unpublished paper, Harvard Graduate School of Education, Cambridge, MA.

Perfetti, C. A. (1985). Reading ability. New York: Oxford University Press.

Popp, H. M. 1975). Current practices in the teaching of beginning reading. In J.B. Carrol & J.S. Chall (Eds.). Toward a literate society. New York: McGraw-Hill.

Smith, F. (1979). Reading without nonsense. New York: Teachers College Press.

Smith, L. C. (1979). An evaluation of studies of long term effects of remedial reading programs. Unpublished doctoral dissertation, Harvard Graduate School of Education, Cambridge, MA.

Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. Journal of Educational Psychology, 8, 323-332.

Thorndike, R. L. (1973). Reading comprehension in fifteen countries. New York: John Wiley and Sons.

What works. (1986). Washington, DC: U.S. Government Printing Office.

Williams, J. P. (1986). Assessment at age seven. Paper commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report, The Nation's report card, August 1986.

The NAEP/ETS Report on the 1986 Reading Data Anomaly: A Technical Critique

Larry V. Hedges
University of Chicago

The NAEP/ETS report on the 1985-1986 reading anomaly by Beaton et al. does a very credible job of proposing hypotheses about the causes of the anomaly and exploring the evidence that these causes might have produced the anomaly. This paper is a technical critique of that report. I will not attempt to recapitulate all of the arguments given in the NAEP/ETS report, but instead will focus on the overall logic of the report, the technical adequacy of procedures used, the link between the technical findings and the conclusions, and avenues that NAEP staff were unable to pursue because of a lack of currently available data.

Nature of the Reading Anomaly

The results of the 1986 reading assessment are regarded as anomalous because the estimated reading proficiency scores at ages 9 and 17 are lower than those estimated by the 1984 report (while reading proficiency scores of 13-year-olds are slightly higher in 1986 than in 1984). Beaton et al. note that " he apparent declines in reading proficiency at age 9 and especially at age 17 are so large during the 2-year period that we doubt that actual changes of this magnitude would have been unnoticed by observers of American education" (page 1). The declines in scaled scores at ages 9 and 17 are about 3 percent of the 1984 values. Given that the standard error of these scaled score estimates is less than 1.1, these shifts are highly statistically significant. If they reflect population values, these declines would represent an enormous shift in such a short period of time. The shift in scaled scores is mirrored by a decline of about 3.6 percent for 9-year-olds and 3.3 percent for 17-year-olds in overall percent correct on reading items and a similar decline in percent correct for embedded sets of reading trend items used in both the 1984 and 1986 assessments.

Changes in variability and distribution shape

Although the Beaton et al. report primarily addresses the decline in means between 1984 and 1986, the change in dispersion is at least as striking. The standard deviation of reading proficiency scores for 9- and 13-year-olds increased by about 10 percent between 1984 and 1986, and the standard deviation of reading proficiency scores for 17-year-olds increased by 25 percent during this period. The change in distribution shape appears to be more complex than a simple change in variance. The upper tails of the score distributions for 1984 and 1986 are quite similar, but the lower tails of the score distributions are heavier in

1986 than in 1984. This suggests a shift of score mass fro near the median of the distribution to th lower tail. If these changes reflect a shift in population values the, would also have enormous educational significance. Consequently, both the shift in mean and in distribution shape appear to be anomalous and an adequate explanation for the anomaly must address both of these shifts. In particular, the explanation must involve effects that interact with level of attainment since the anomaly has apparently led to a change in distribution shape.

General Criticisms of the Report

I have four general criticisms of the assumptions that were made in both the choosing and interpreting analyses of the anomaly in the report by Beaton et al. The first assumption, as stated before, was that the anomalous means were of primary interest. This led to an almost exclusive focus on the decline in means, and changes in the shape of the distribution were largely ignored. Although the sampling design makes analyses of dispersion somewhat more difficult than analyses based on means, such analyses are crucial to a complete understanding of the anomaly. The recognition of requirement that a complete explanation for the anomaly must include effects that interact with level of attainment might have stimulated alternative explanations or helped to rule out others.

The second assumption underlying the Beaton et al. report is that the anomaly is the result of a single effect. Each of the potential explanations for the anomaly was examined in turn to see whether it alone could produce an effect large enough to account for the anomaly. If the probable effect was not large enough to completely account for the anomaly, the explanation was dismissed. This seems particularly dangerous because several of the potential explanations can account for substantial fractions of the observed change in means. That is, each could potentially produce a change in mean scores that would be considered large in an absolute sense, albeit not as large as the anomaly. For example, the effect of a shift in modal grade of respondents in the 9-year-old sample is estimated as producing an effect that could be as large as 50 percent of the observed anomaly. Surely this effect is worth noting because of its absolute magnitude and because it could well be a major contributor to the anomaly at the 9-year-old level. Given the many changes in the design, implementation, and analysis of the NAEP between 1984 and 1986, it may be unrealistic to expect the anomaly to be the result of any single effect. It seems more likely that the anomaly may be the result of several causes.

The third assumption is that the best measure of the size of the anomaly is the comparison between the results in 1984 and those in 1986. The assertion is that 1986 results are anomalous because they differ from 1984 results by more than would be expected based on trends from previous NAEP assessments or based on contemporary trends in other reading achievement data. The possibility remains that the 1986 reading changes were made in the 1984 assessment. Anomalous results during the 1984

assessment that were not previously detected could have contributed to the so-called 1986 reading anomaly. For example, the 17-year-old reading scores in 1984 were higher than expected. If the 1984 results were simply extrapolated from the 1971-1980 linear trend, more than 25 percent of the 1986 anomaly would disappear.

The Beaton et al. report considered the possibility that the 1984 results were themselves anomalous, but rejected it because anomalous results in 1984 could not by themselves have explained the apparent 1984 to 1986 reading anomaly. It seems unwise to reject a priori the possibility that anomalies in 1984 reading scores contributed to the 1986 reading anomaly, particularly at age 17.

The fourth assumption was unstated but might be regarded as implied by the use of 95 percent confidence intervals for NAEP assessment results in Figure 2.1 of the Beaton et al. report. The use of cross-sectional standard errors suggests (at least to less sophisticated readers) that a reasonable measure of the expected stochastic variation between years (i.e., the standard error of the difference between mean values at different assessments) is

$$\sqrt{SE_{1984}^2 + SE_{1986}^2}$$

or about 1.4 SE_{1986} if we assume that standard errors for both years are about equal. This is a potentially misleading assumption because there are many components of between (1984 and 1986) assessment variance that are not contained in cross-sectional standard deviations. In fact, the Beaton et al. report identifies several changes in the respondents or in the assessment instrument that appear to be associated with effects that are larger than one cross-sectional standard error.

Hypotheses about Reasons for the Score Decline

The report by Beaton et al. considered seven general classes of potential explanations for the score decline: shifts in population, unrepresentativeness of sample, changes in the measuring instrument, changes in administrative procedures, failures of quality control, artifacts of scaling, effects of a subset of items or of item response pattern and artifacts of booklets and blocks. They also considered two miscellaneous hypotheses concerning the effects of external events and the possibility that 1984 scores were unusually high. The examination of each of the hypotheses was designed to reveal if that hypothesis could by itself explain the decline in mean score between 1984 and 1986.

Population and Sample Hypotheses

The hypotheses that populations have shifted or that the samples are non-representative are among the most obvious explanations for the anomaly, and they were investigated thoroughly. There seems little to believe that the anomaly was caused by errors in weighting procedures.

Similarly, the fact that declines in percent correct occurred in virtually all demographic sub-groups provides convincing evidence that the anomaly was not primarily a result of a purely demographic shift in population sizes or response patterns. There were, however, hypotheses related to population and sample that appear to explain some portion of the anomaly. These are discussed individually below:

Date of the Data Collection. The 9- and 17-year-old samples were assessed earlier in 1986 than in 1984 and the 13-year-old sample was assessed later in 1986 than in 1984. The average difference between 1986 and 1984 in date of assessment corresponded to -22 days for 9-year-olds, +4 days for 13-year-olds, and -18 days for 17-year-olds. It is interesting to note that these differences in date of assessment correlate .996 with the difference in percent correct (-3.6, +.8, and -3.3 respectively) on the reading trend items for these three age groups. Beaton et al. conclude that these relatively small differences in date of testing have only a small effect on attainment ("at most one scale score point," page 33). However, the logic of their argument is at least debatable. First, their analysis is based on the idea that attainment is a function of age, not time in school. They obtained their estimate of effect by linear interpolation of 1984 NAEP results by regressing scale score on age. It is perhaps more reasonable to argue that growth in attainment is better modeled as a function of time in school and that the function is not linear. For example, there is some empirical evidence that attainment actually declines over the summer when students are out of school, and there is certainly anecdotal evidence that little increase in attainment occurs during the first and last few weeks of school and during the week just before winter and spring vacations. This suggests that not all school days are equal. Moreover, the particular school days that were experienced by the 1984 sample but not by the 1986 sample were likely to be among the most productive in raising attainment. Thus linear interpolation over the entire year substantially underestimated the effect of difference in date of testing--perhaps by 100 percent but probably by less than 200 percent. It seems unlikely that time of testing alone could account for the anomaly, but it might very well account for 2.0 to 2.5 scale score points.

Another possible effect of time of testing is noteworthy. If, as seems reasonable, the function relating time in school and attainment is different at different levels of attainment, the difference in date of testing could very well lead to a greater number of lower scoring 9- and 17-year-olds in 1986 than in 1984. For example, if students with low attainment actually grew at a relatively faster rate in February and early March, then the fact that the 9- and 17-year-olds in 1986 assessment did not have this time in school would have differentially increased the number of low scorers. Although the NAEP program cannot directly provide information about rates of growth in attainment, such information from any source would be useful to help understand the effects of assessment schedules on expected scores and on diversity.

Attributes of low scoring students. The analyses that investigated attributes of low scoring students provided rather convincing evidence

that the score decline for 17-year-olds was not restricted to one or a few demographic groups. The analysis designed to determine if a few schools are implicated in the score decline for 17-year-olds does not suggest any obvious pattern of concentration. It involves an examination of the frequency distributions of 1984 and of 1986 school level scores on the sample group of items. It is interesting to note, however, that the only intact set of items used to search for school effects (apparently the only set of items contained within the same block in both years) were contained in block R4. This is the same block that was identified subsequently as producing somewhat different percent correct (up to 5 percent different) depending on the position of block R4 within the booklet. The implications of the susceptibility of this booklet to context (within booklet) effects substantially limits the school effects study or vice versa.

The study of school effects was also limited in that it did not address the issue of within-school variation in scores. Variations across schools in administrative procedures could have resulted in different within-school variance components. A more thorough analysis of variance components would have been useful in understanding the contribution of variations in administrative procedures to the anomaly.

Measuring Instrument Hypotheses

Although the changes suggested by the reading anomaly would be very large if there were population changes, they do not reflect particularly large changes in the individual level performance. For example, the change of 3.3 in percent correct for the 39 reading trend items at the 17-year-old level corresponds to each student answering approximately 1.25 fewer items correct in 1986 than in 1984, and the change of 3.6 in percent correct for the 30 reading trend items at the 9-year-old level corresponds to only 1.08 items. Such small effects might plausibly be the result of the several changes in format and administration of the assessment instrument. Unfortunately, very little empirical evidence is presented to help assess the magnitude of the effects of such changes on assessment results. In the absence of such information one can only speculate, as I have done below. (Note here I would like to add something from studies of similar effects.)

Booklet Format and Scoring. Two changes in booklet format and scoring seem particularly suspicious. The 1986 assessment used "fill in the oval" format for responses and machine scoring, while the 1984 assessment used "circle the letter" format and key entry of responses. Since the new response format and scoring method are in some ways more demanding and less forgiving of errors in procedure, it seems plausible that these changes might lead to a smaller proportion of responses coded as correct. This may explain part or all of the score decline in the 9- and 17-year-old samples. It does not seem reasonable that such effects might be more pronounced among students with low attainment, which might help to explain the larger number of low-scoring students in 1986. Although differences in booklet format and scoring are likely to be contributors to the anomaly, it is puzzling that no large differences appear at age 13, where

the same format and scoring were used. There is some indication, however (see Table 1), that positive effects of other artifacts (e.g., changing patterns of nonresponse and scaling) could be masking a negative effect of format and scoring. The experiments embedded in the 1988 assessment designed to investigate format and scoring will provide estimates of these effects.

Administrative Change Hypothesis

Administrative changes between 1984 and 1986 might also have been responsible for part of the anomaly, but there is no empirical evidence about the magnitude of the effects that might have resulted from administrative changes. Moreover, although some of the changes apply only to 17-year-olds, none apply only to 9- and 17-year-olds and, therefore, none would explain why the scores of 9-year-olds decreased but those of 13-year-olds did not. There appears, however, to be several other possible explanations for the anomaly at age 9 (see Table 1). The increase from 20 to 35 of the average number of individuals assessed in each session may have contributed to the anomaly for the 17-year-olds. Similarly, the introduction to the 17-year-old students of up to 5 teachers during the assessment session may have contributed to the decline. Either of these effects might plausibly had a greater effect on low-scoring students, but there is no clear evidence to this effect.

Quality Control Hypotheses

The anomaly corresponds to only a few percent and consequently even a source with a small rate of errors could contribute a substantial proportion of the total anomaly. The studies of quality control suggest that the data entry process was very accurate. The estimated error rate would contribute very little to the anomaly. However, a large sample is necessary to convincingly search for small effects. It is possible (although it seems unlikely) that there are clusters of booklets with higher error rates than those uncovered in the studies of quality control. For example, among the 2.3 percent of the damaged or irregular student booklets that were keyed by hand it might have been desirable to select a sample of booklets stratified by keypuncher (rather than a simple random sample) to assure that there were not important differences across keypunchers in accuracy.

Scaling Hypotheses

Because NAEP uses a complex scaling procedure, the possibility that the anomaly was an artifact of scaling procedures was investigated. The scaling hypothesis is partially disconfirmed by the raw data (percent correct) on the sets of reading trend items that were identical in both the 1984 and the 1986 assessments. The fact that there was a decline in percent correct both overall and for reading trend items in the 9- and 17-year-old samples that roughly corresponds to that of scale scores suggests that scaling alone cannot explain the anomaly. Using an approximate group level IRT model (Mislevy, 1983) and making certain simplifying assumptions, Beaton et al. estimated from the mean percent of

correct responses the change in scaled score between 1984 and 1986. In each the estimated changes were smaller in absolute magnitude than the changes computed using the usual scaling method, but in no case was the difference between the two scaling methods greater than 2 scale score points. This analysis confirms that the anomaly is not an artifact of the scaling process, although the scaling process may have tended to magnify a real difference in raw scores.

Item Level Hypotheses

The NAEP data collection recognizes five different types of responses for each item: correct answers, incorrect answers, "I don't know" (IDK), omits (when a student responds to an item later in the block), and not reached. Because IDK and omit responses are treated differently in scaling than in computing percent correct for each item, changing patterns of these nonresponses over time can have effects on the scores estimated in the assessment. This is because changing response patterns (or nonresponse patterns) actually imply changes in the population providing data for the scoring process. Consequently, changes in patterns of responses could lead to both artificial changes in mean assessment scores and to changes in distribution shape.

The Beaton et al. analysis of the nonresponse patterns in 1984 and in 1986 suggests that changing patterns of nonresponse could be responsible for a part of the anomaly. Using Mislevy's group level IRT approximation, if the 17-year-olds in 1986 had exhibited the same pattern of nonresponse as those in 1984, the difference would have been reduced by about 20 percent or 2 scale points. Similar calculations for the 9-year-olds suggest that changing patterns of nonresponse could account for about 33 percent of the decline (about 2 scale score points). Calculations for the 13-year-olds suggest that changing patterns of nonresponse could actually account for an increase of about 1.5 scale points in 1986.

Booklet and Block Hypothesis

In 1984, each NAEP assessment booklet containing reading items consisted of one of three blocks of reading items and one to three blocks of writing items to yield a total of three blocks per booklet. In 1986, each NAEP assessment booklet was also divided into three blocks of items. Each block consisted of items in the same content area, but the non-reading blocks included items in the content areas of mathematics, science, computer competence, history, and literature. Thus one difference between the 1984 and 1986 assessments is the grouping in 1986 of reading and non-reading items into the same booklets. The analyses designed to examine booklet and block effects were not very exhaustive. An analysis of percent correct as a function of block position within booklet and subject matter of items preceding the reading items suggests that one reading block (R4) is particularly susceptible to position effects, giving the lowest percent correct when it was preceded by two non-reading blocks. The potential effect of this block was examined by recomputing 1986 scores by eliminating individuals who received assessment booklets where reading block R4 was preceded by two non-reading blocks.

This eliminated only about 9 percent of the decline in overall average percent correct.

The analyses conducted do not convincingly rule out the possibility that combining blocks of reading items in the same booklet with blocks of items from other subject matters contributed to the anomaly, however. An examination of Table 11-2 in Beaton et al. suggests that reading scores are typically lower when a reading block follows another content area than when a reading block follows another reading block. The difference between percent correct when reading follows reading (as opposed to another content area) is one crude estimate of the effect of grouping reading with other subject matters. A crude analysis combining these effects across blocks and weighting reading block R4 as much as all other reading blocks combined (because it was used as often as all other reading blocks combined) suggests that the effect of combining reading with other subject matters could be as large as 1.5 percentage points - nearly half the size of the anomaly.

Combining Effects of Potential Explanations for the Anomaly

The Beaton et al. report concluded that they had failed to explain the anomaly because no single source seemed likely to produce effects as large as the anomaly. However, as Table 1 illustrates, several sources produced effects that were estimated to be substantial fractions of the observed anomaly. Some of these effects are two to three cross-sectional standard errors. If their effects are approximately additive, several of these sources could jointly explain most or all of the anomaly. For example, shift in the proportion of modal grade respondents and changing pattern of nonresponse could together account for 83 percent of the anomaly at age 9. The effects of date of assessment could account for the rest. At age 17 the effects of date of assessment, changing patterns of nonresponse, and scaling effects could account for nearly 60 percent of the anomaly if effects were additive. The effect of mixing reading blocks with other subject matters might be as large as 50 percent of the anomaly if the crude analyses outlined in the previous section are correct. It is easy to imagine that the effects of changes in booklet format and scoring and administrative changes are of the same magnitude as some of the other effects in Table 1. If so, these sources could together produce effects as large as those observed as the anomaly.

Table 1

Estimated Effects on 1986 Scaled Scores of Sources Likely to Produce Largest Contributions to the Anomaly

Source	Age 9		Age 13	Age 17	
	Points	Percent of anomaly	Points	Points	Percent of anomaly
Date of assessment - linear trend	-1	-17	+	-1	-9
(Date of assessment - S-shaped trend)	(-2)	(-33)	(+)	(-2)	(-19)
Shift in modal grade	-3	-50		+1	+9
Scaling	-.2	- 3	+1.5	-2	-19
Changing pattern of nonresponse	-2	-33	+1 5	-2	-19
Reading block effects	?	?	?	-1	9
(Mixing reading with non-reading, non-writing content)	?	?	?	(-5?)	(50)
Booklet format and scoring	?	?	?	?	?
Administrative changes	?	?	?	?	?

Note: Values reflect the amount by which 1986 scores were changed due to a source. Thus negative values reflect amounts by which the source accounts for the anomaly at ages 9 and 17.

Note also that the effects of several sources at grade 9 suggest that 1986 scores could be underestimated by over 3 scale score points. This suggests the possibility that the 1986 results for 13-year-olds are unchanged from 1984 because the positive effects of artifacts such as changing patterns of nonresponse were cancelled by the negative effects of other artifacts such as booklet format and scoring. Obviously this is speculative, and firm conclusions must await the results of the experiments embedded in the 1988 assessment.

References

Beaton, A. E. (1987). Implementing the new design: The NAEP 1983-84 technical report. (NAEP Report 15-R-01). Princeton, NJ: Educational Testing Service.

Beaton, A. E., Ferris, J. J., Johnson, E. G., Johnson, J. R., Mislevy, R. J., & Zwick, R. (1987). The NAEP 1985-86 reading anomaly: A technical report. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1983). Item response models for grouped data. Journal of Educational Statistics, 8, 271-288.

Reading Trend Data from the National Assessment of Educational Progress: An Evaluation

Janet Baldwin
American Council on Education

In evaluating trend data from the National Assessment of Educational Progress (NAEP), an important criterion for judging the accuracy of score interpretation over time is that the scores from each assessment have consistent meaning from one test administration to the next. The purpose of this paper is to examine the accuracy of NAEP reading trend data by 1) describing trends in the reading skill performance of 17-year-old high school students as measured by the NAEP in 1971, 1975, 1980, 1984, and 1986[1]; 2) evaluating the consistency of test content over time; and 3) examining the influence of changes in test content and in test development and administration procedures on the accuracy of NAEP trend data. Finally, recommendations are made for improving the accuracy of trend data from the NAEP.

Trends in Reading Skills Measured by NAEP

The NAEP was designed to furnish information regarding the educational achievements of students to all those interested in American education, "indicating both the progress we are making and the problems we face" (NAEP, 1970; p.1). According to NAEP's 1970 Reading Objectives, the purpose of the assessment is to provide helpful information about the progress of education that is understandable to laymen as well as professional educators. To accomplish this purpose, "some new procedures were followed in constructing the assessment instruments that are not commonly employed in test building" (NAEP, 1970; p.2). One of these new procedures, a consensus approach to the development of content objectives, apparently permitted changes to be made in the aims of each assessment as well as in the procedures applied.

Most national testing programs have in place procedures for assuring comparability over time in test meaning and score interpretation. The most direct way that scores on a test may be made comparable over time is by administering the same test in the same way to samples of examinees drawn from the same population. When this is not feasible or desirable, scores from different tests may be equated to achieve comparability, or comparability may be built into the test through field testing and evaluation of items, tests, and procedures and through consistent test construction and administration practices. Until recently, NAEP assessments do not appear to have followed these approaches for assuring comparability over time.

In 1984, the method of reporting trend data from the NAEP was changed from the proportion of students answering each item correctly, or percent correct scores, to estimated true scores on a proficiency scale based on a hypothetical reading proficiency test (Beaton, A.E., 1987a; NAEP, 1985). Although the 1984 NAEP reading test was equated with all previous NAEP reading assessments, the results are no longer reported in terms of specific reading exercises, content, or objectives, as in the past, but rather are reported in terms of average reading proficiency scores. That is, the reading achievement measured in the NAEP assessment is no longer dependent on, or linked to, specific exercises but rather "reflects a proficiency in comprehending or constructing meaning from a broad range of prose materials" (NAEP, 1987, p. 23). Because the content objectives on which the proficiency scores were based appear to have changed between 1984 and 1986 (in the substitution of new content categories and by excluding reference skills in 1986), the comparability between assessments administered through 1986 does not appear to have been maintained.

Content of NAEP

The NAEP reports students' aggregate achievement relative to a criterion or level of competence in specified content areas. In the first three NAEP assessments of reading, the criterion was described in terms of average performance on a set of common items averaged over reading content categories, such as literal comprehension, inferential comprehension, and reference skills (NAEP, 1976, 1981). The same criterion was applied in the comparison between the third and fourth assessments, although a different set of common items was used (NAEP, 1985). In comparing the fourth and fifth NAEP reading assessments, the criterion was described in terms of rudimentary, basic, intermediate, adept, and advanced levels of reading achievement relative to the reading proficiency scale (Beaton, 1987a, 1987b; NAEP, 1985).

As is evident from information presented in recent NAEP reports (Applebee, Langer, & Mullis, 1987; Beaton, 1987a; Beaton, 1987b) and in the NAEP Reading Objectives publications (1970, 1974, 1980, 1984, and 1987), the relationship between reported scores and the domain of content objectives appears to have changed over assessments. Therefore, in order to evaluate the trends in what was commonly measured in each assessment, the performance of 17-year-olds on common sets of items administered in the 1971, 1975, 1980, 1984, and 1986 assessments was examined.

Content Comparability

The content meaning of NAEP trend results becomes more clear when student performance on common items is examined by content categories over time. Because the content categories used for the 1971, 1975, 1980, and 1984 assessments included Literal Comprehension (LC), Inferential Comprehension (IC), and Reference Skills (RS), these categories were used here to evaluate the content comparability of NAEP over assessments.

NAEP trend data are presented in Table 1 as mean percent correct scores (percentage of students who answer the item correctly) for three content

categories of reading items administered in multiple assessments under under Paced[2] and Balanced Incomplete Block[3] (BIB) conditions (Beaton, 1987a, 1987b). Under the circumstances of questionable comparability over time, this approach has the advantage of presenting achievement information in a form which is closer to the level of raw data than, say, proficiency scale scores, and provides a point of reference for examining changes in achievement based on content categories which are familiar to most test users. There are limitations in this approach, however. The reported level of performance has meaning only in terms of the particular items included in the comparison and some comparisons include relatively few items. Moreover, the common items may not be representative of the broad range of reading content objectives covered by the complete assessment at each time.

Trend comparisons based on common items. For the first three comparisons (1971 - 1975, 1975 - 1980, and 1980 - 1984), trend items were administered under Paced conditions. The 1984 - 1986 comparison was made under BIB procedures. Because only four items from the first four assessments were included in the 1986 NAEP, these items will not be compared across all five assessments.

For 17-year-olds, a common set of 71 reading exercises administered in each of the assessments for 1971, 1975, and 1980 was used to report trends in reading achievement. Of these original 71 trend items, only 19 were administered in 1984. Therefore, 71 exercises are common to the first three assessments and 19 are common to the first four assessments. For the 1980-1984 comparison, in addition to the 19 items common to the earlier assessments, 34 new items were included, bringing the total of common items in that comparison to 53. The comparison between items administered under 1984 Paced conditions and 1984 BIB conditions is based on 17 common items which also were common to the 1980 assessment. The 1984-1986 (BIB) comparison is based on these same 17 common items. Therefore, 17 items are common to the 1980 (Paced), 1984 (Paced), 1984 (BIB), and 1986 (BIB) assessments.

The results from analyses based on items common to two, three, or four assessments are presented in Table 1, below. It is important to note that the relevant information here is not the magnitude of the mean percent correct scores, as this varies depending on the difficulty of the items included in the analysis, but rather the relative changes in mean percent correct from one year to the next. For a given set of common items, the table should be read by rows across years.

As shown in Table 1, trends in reading achievement are based on different sets of common items for LC, IC, RS, and Other. The 19 common items administered in 1971, 1975, 1980, and 1984 indicate a steady increase in total reading achievement for 17-year-olds during this period, from 69.2 to 69.9. However, the trends in LC and IC, based on 13 and 6 items respectively, suggest the increase in this total score performance through 1980 may have been due to improvements in LC, as performance on IC items during this time declined. From 1980 to 1984. however, performance on IC increased while performance on LC leveled. Although generalizations based

on so few items may not be reliable, these results do illustrate that trends in component scores (i.e., LC and IC) provide different kinds of information from trends in total scores. Because items measuring reference skills were omitted from NAEP trend comparisons in 1986, additional trend comparisons over multiple assessments are given in Table 1 based on common items which exclude items measuring reference skills. For this reason, totals in the table are based on different sets of common items, some including only LC and IC, and others including RS and Other. Although totals based on LC and IC items only indicate a steady, though slightly increasing, trend from 1971 to 1984, totals from 1984 to 1986 declined by nearly 4 percentage points, from 67.7 to 63.8. NAEP's reported reading scores for 17-year-olds in 1980 and 1984, which indicated an increase in average p-values fr 73.4 to 74.5 (NAEP, 1985), are based on items categorized as LC, IC, RS, and Other. When trend comparisons for 1980 and 1984 are examined by separate content areas, however, it becomes apparent that the reported increase in reading performance in 1984 is due primarily to the considerable improvement in performance on the RS and Other items. The average p-values for each of these two content categories increased by about 3 percentage points from 1980 to 1984.

The comparison between the Paced and BIB conditions in 1984 indicates that the introduction of BIB procedures depressed the p-values of the 17 common items by about 2.5 percent and the effect was relatively consistent across content categories. The comparison between BIB 1984 and BIB 1986 shows marked declines in performance on the common items with the greatest impact on the literal comprehension category, which decreased from 70.6 to 65.2, and on the three items included in the category labeled Other, which decreased from 79.8 to 73.6.

Table 2 presents the proportion of items in the content categories LC, IC, RS, and Other for assessments administered in 1971, 1975, 1980, and 1984 and 1986. In comparing the content coverage represented by adjacent NAEP assessments, it is notable that the 1971, 1975, and 1980 assessments were based on the same number and proportions of items in each content area. For the 1980-1984 comparison, the number and proportion of common items in each content area differed from those in previous comparisons. Although the content categories themselves changed completely in 1986, the 1986 proportions presented in Table 2, for the purpose of comparison, are based on the LC, IC, and Other categories for the 17 items common to the 1984 and 1986 assessments.

As shown in Table 2, of the 71 items common for the 17-year-olds assessed in 1971, 1975 and 1980, .49 measured LC, .35 measured IC, and .15 measured RS. Of the 53 common items from 1980 and 1984 assessments, .38 measured LC, .23 measured IC, and .34 measured RS. The category Other comprised .06 of the total. The doubling of the proportion of RS items and the corresponding decrease in proportions of LC and IC from 1980 to 1984 seriously distort the comparability of the total scores produced in those assessments. Of the 17 common items from 1984 to 1986, .41 measured LC, .41 measured IC, and .18 measured Other, reflecting yet another shift in emphasis among the content categories. In 1986, the item classifications were changed from LC, IC, and RS to categories labeled Deriving Information

(DI), Integrating and Applying Information (IAI), and Evaluating and Reacting (ER). Items previously labeled RS were omitted from 1986 trend analyses and some trend items previously labeled LC and IC became either DI or IAI. The items previously labeled Other became either ER or IAI (NAEP, 1988). While the new content classifications may very well represent improvements, their comparability with previous classifications appears to be lost.

Confounding Influences on Comparability of NAEP Trend Data. The 1984 and 1986 assessments not only introduced changes in scaling and reporting, but also in booklet construction and test administration procedures. Moreover, the procedures and materials used in 1986 were considerably more complex than those used in 1984, especially for 17-year-olds.

For the 9-year-olds, the 13-year-olds, and the 17-year-olds, NAEP booklet content was more diverse (including assessments of mathematics, science, and computer applications) and administrative procedures were more varied (combining in the same testing session both tape-recorded and student-read instructions). For 17-year-olds in 1986, the Teacher Questionnaire administrative procedures prior to the administration of the NAEP exercises were far more complex and potentially more distracting to students than in 1984. In addition, the number of students per testing session increased by 75 percent in 1986 and for a substantial portion of the 17-year-olds, the 1986 administration included an assessment of literature and history. Because such changes can influence student performance and item difficulty, the comparability of the proficiency scale scores based on even the same items administered under such differing procedures may be questionable.

Evidence of possible context effects on items due to booklet content is provided in Table 11-2 of the 1987 Technical Report (Beaton, 1987b). This table presents the average percent correct for the 1986 reading blocks administered to 17-year-olds when the items were in positions 1, 2, or 3, following blocks of reading or other content. The average percent correct for reading block R4 when located in the first position in the booklet, 72.1, declines by 2.5 percent when located following one block of other content. It declines by 4.8 percent when located following two blocks of other content. Oddly, an unusually large proportion of the sample (33 percent) was administered booklets with reading block R4 in positions 2 and 3, following one or two blocks of other content. The nature of the decline in performance of 17-year-olds on ten of the items in reading block R4 is especially notable when the frequency distributions of these items in the 1984 and 1986 administrations (Figure 5-1; Beaton, 1987b) are compared to the distributions of total reading proficiency scores for 17-year-olds in those years. The shapes of these distributions are quite similar.

Conclusions

During the past two decades, many changes have been made in the test content and in the test development, administration, and scoring procedures for the NAEP. Not only has each NAEP been designed to reflect educational

practices currently in vogue, but the set of exercises used to provide trend information have varied over the years. These changes in procedures and in test content confound the meaning and interpretability of the NAEP trend data.

In general, the portion of the assessment on which trend data are based must be shown to measure the same content objectives in all relevant subpopulations over time, be administered using the same procedures, and interpreted in reference to a clearly defined domain of content or behaviors. Future plans for NAEP assessments should address the need for consistency in assessment and content objectives, booklet content, and test development and administration procedures in order to ensure comparability in trend data over time.

Recommendations

The following recommendations suggest ways to improve the consistency in the meaning and interpretability of NAEP scores over time.

a. Identify a central core of important instructional objectives for the Nation on which trends in reading achievement will be reported. Although a consensus approach for defining objectives has been followed for each assessment in the past, little attention appears to have been paid to the continuity of consensus over time. The core components of the domain should be those widely and commonly judged to represent the most important and pervasive skills, knowledge, and developed abilities on which information is required over time. The criteria for selecting objectives within each component of the domain should include their stability and usefulness over time. When these objectives are selected, they should remain constant over time.

b. Define the domain of performance, or behavior, of practical importance to which inferences from test scores will be made. Although the description of the reading proficiency scale addresses this issue, the scale reduces a complex set of skills to a single score and the usefulness of this approach to practitioners and to policy makers is yet to be demonstrated. A well-defined domain of reading behaviors should guide the development of items used in the assessment. The NAEP exercises currently available were developed over the years to measure various objectives, none of which, until recently, included the assessment of reading proficiency as defined by the current scale. The definition of this domain should be further refined and the definition should remain stable over time.

c. Specify systematic methods by which exercises will be developed. These methods should be designed to produce exercises which, within each component of the domain, are interchangeable. In this way, exercises may be selected from a pool of comparable exercises within each component of the domain and comparable test forms of known difficulty may be constructed. Such methods also will ensure that future exercises written by different writers will be functionally equivalent to those used on previous forms.

d. Develop a set of test specifications for the portion of the assessment on which trend results are based. These specifications should indicate the relative emphasis to be given to each component of the domain. For the trend portion of the assessment, these emphases should remain constant over time. Exercises measuring various content areas and level of cognitive operations should be represented in proportion to their relative importance. By maintaining the same emphasis in content and cognitive processes in each assessment, stability of the assessment instruments and assessment results will be increased.

e. Field test all item formats used in the NAEP for each age group to determine if the format is feasible for the intended examinee group. Some formats appropriate for 13-year-olds, for example, may not measure skills in the same way for 9- and 17-year-olds, and vice versa. Because there is considerable variability in the length of reading passages (from 30 to 2000 words), the appropriateness of passages of varying lengths should also be tested for different age groups. NAEP exercises should be evaluated for statistical and substantive adequacy prior to inclusion in operational forms of the assessment. The test developer should specify the procedures followed and criteria applied in such evaluations. When new items are needed for trend purposes, they should be selected on the basis of their functional and statistical equivalence to those items they are replacing. Evaluation strategies should include item analysis procedures which are appropriate for criterion-referenced or objectives-referenced assessment purposes.

---

[1]For convenience, the 1970-71, 1974-75, 1979-80, 1983-84, and 1985-86 assessments are identified as 1971, 1975, 1980, 1984, and 1986 respectively.

[2]Paced administration provided tape-recorded instructions to students and progress through the assessment is paced by audio-tape. The same package of exercises is administered to all students within a session (Beaton, Johnson, and Ferris, 1987).

[3]Balanced Incomplete Block (BIB) administration is a complex variant of multiple matrix sampling which divided up the total assessment time into small blocks. Students in an assessment session are given different booklets containing different blocks of exercises. Students have a specific block of time within which to complete a booklet (Beaton, Johnson, and Ferris, 1987).

Table 1. National Mean Percent Correct for 17-Year-Olds in Five Reading Assessments Based on Different Sets of Common Items

Content Area	No. of Items	Paced 1971	Paced 1975	Paced 1980	Paced 1984	BIB 1984	BIB 1986
Literal	35	72.2	72.7	72.0			
Comprehension (LC)	13	71.3	72.5	72.9	73.0		
	20			76.2	75.5		
	7			73.6	73.2	70.6	65.2
Inferential	25	64.2	63.3	62.1			
Comprehension (IC)	6	64.7	62.6	62.2	63.2		
	12			70.2	71.4		
	7			67.1	67.4	64.9	62.5
Reference Skills (RS)	11	69.4	70.1	70.2			
	18			71.2	74.0		
Other (O)	3			79.2	82.2	79.8	73.6
TOTALS (LC+IC+RS)	71	68.9	69.0	68.2			
(LC+IC)	19	69.2	69.3	69.5	69.9		
(LC+IC+RS+O)	53			73.4	74.5		
(LC+IC+O)	35			74.4	74.7		
(LC+IC)	32			74.0	74.0		
(LC+IC+O)	17			71.9	72.4	69.9	65.5
(LC+IC)	14			70.3	70.3	67.7	63.8

Source: Reading Report Card (1985). Calculations by author based on unpublished data from NAEP, Educational Testing Service, Princeton, N.J. Mean percent correct scores for each item were averaged within content categories. Janet Johnson and Kentaro Yamamoto, personal communications.

Table 2. Proportions of Trend Items in Three Content Categories Administered to 17-Year-Olds in Five Reading Assessments

	(No. of Items)	1971	1975	1980	1984	1986
Literal Comprehension	(35)	.49	.49	.49		
	(20)			.38	.38	
	(7)				.41	.41
Inferential Comprehension	(25)	.35	.35	.35		
	(12)			.23	.23	
	(7)				.41	.41
Reference Skills	(11)	.15	.15	.15		
	(18)			.34	.34	
Other[1]	(3)			.06	.06	
	(7)				.18	.18
Total[2]	(71)	.99	.99	.99		
	(53)			1.01	1.01	
	(17)				1.00	1.00

1 Total scores reported in 1980 and 1984 included three items categorized as "Other". However, these items were not reported separately and do not appear to have been included with any other content category.

2 Totals do not sum to 1.00 due to rounding error.

References

Applebee, A. N., Langer, J. A., & Mullis, I. V. (1987). Who reads best? Factors related to reading achievement in grades 3, 7, and 11. Princeton, NJ: Educational Testing Service.

Beaton, A. E., Ferris, J. J., & Johnson, E. G. (1987). The assignment of exercises to students. In Implementing the new design: The NAEP 1983-84 technical report. Princeton, NJ: Educational Testing Service.

Beaton, A. E., Ferris, J. J., & Johnson, E. G. (1987). The NAEP reading scale. In Implementing the new design: The NAEP 1983-84 technical report. Princeton, NJ: Educational Testing Service.

Beaton, A. E., Ferris, J. J., Johnson, E. G., Johnson, J. R., Mislevy, R. J., & Zwick, R. (1987). The NAEP 1985-86 reading anomaly: A technical report. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1970). Reading objectives. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1974). Reading objectives: Second assessment. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1976). Reading in America: A perspective on two assessments. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1980). Reading and literature objectives: 1979-1980 assessment. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1981). Three national assessments of reading: Changes in performance, 1970-80. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1984). Reading objectives: 1983-84 assessment. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1985). The reading report card: Progress toward excellence in our schools, Trends in reading over four national assessments, 1971-1984. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1987). Reading objectives: 1986 and 1988 assessments. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (February 1988). Year 17 Reading Item Classifications. Internal NAEP document. Princeton, NJ: Educational Testing Service.

Mathematics Trends in NAEP: A Comparison with Other Data Sources

Tej Pandey
California Assessment Program

Since 1969, NAEP has profiled achievement of the Nation's 9-, 13-, and 17-year-olds attending public and private schools in certain subject areas, including mathematics. NAEP has conducted four assessments in mathematics: in 1972-73, in 1977-78, in 1981-82, and in 1985-86. This paper is a compendium of papers commissioned by the National Center for Education Statistics (NCES), Department of Education to investigate the various aspects of NAEP. This paper examines, in particular, the accuracy of mathematics trends in NAEP by comparing trends from other major databases. The paper also examines the structure and quality of mathematics items used for establishing trends. Finally, conclusions and recommendations are offered for strengthening future assessments, especially in light of NAEP's expansion to make State-by-State comparisons.

Background

NAEP is a unique assessment system in the Nation mandated by Congress to assess the knowledge, skills, understandings, and attitudes of young Americans. One of the important outcomes of NAEP assessment is the achievement trend providing growth in students learning. However, recently some concerns have been raised about the accuracy of NAEP trends, because the last assessment in reading showed a precipitous decline from 1984 to 1986. The decline was so large that Beaton et al. (1987) noted, "The apparent declines in reading proficiency [illegible]ge 9 and especially at age 17 are so large during the 2-year period that we doubt that actual changes of this magnitude would have been unnoticed by observers of American education" (p. 1). Since NAEP is continually incorporating modern technical improvements into its procedures, it is natural to ask the question whether the anomalous reading results for 9- and 17-year-olds are due to changes in methodological and/or administrative procedures or if they represent "true" changes in the achievements of 9- and 17-year-olds. Furthermore, if the reading results are an anomaly, questions arise about the adequacy of trend data for other subjects assessed.

Some Significant Changes in NAEP

The two important components that can skew trend data are the cycle-to-cycle changes in questions used for gaining trend information and the methodological and administrative procedures to collect the data. Each NAEP assessment contained a range of questions on a set of objectives developed by nationally representative panels of mathematics specialists, educators, and concerned citizens. NAEP uses a small set of unreleased exercises constant throughout various cycles in order to anchor the results across time. With each successive assessment, the objectives are based on

the framework used for the previous assessment, with some revisions that reflect current changes and trends in school mathematics. Table 1 shows the number of questions used in various scales for establishing trends from 1978 to 1986.

Related to methodological and administrative changes, NAEP has used a nonoverlapping item sampling design for allocating items to test forms in the 1974 and 1978 assessments. In 1982, NAEP opted to use a Balanced Incomplete Block (BIB) design for item allocation. In 1982, blocks for each subject, such as reading and mathematics, were administered separately, whereas in the 1986 assessment two or more blocks from the subject areas of reading, mathematics, science, and computer literacy were combined. This change in design--combining blocks from various subject areas--necessitated changes in test administration procedures. For instance, in 1986 NAEP relaxed pacing in test administration by discontinuing use of prerecorded audio tapes used to pace students. In 1986, some changes were also made in the dates on which students were tested.

NAEP has also changed the format and statistics used for reporting the results. Prior to 1985-86, results were reported in percent correct units on an exercise-by-exercise basis and for aggregate of exercises; now, however, NAEP is using item response theoretic models to report results across years as well as across age levels on a common content-referenced scale.

Nature of Investigations for Trends

Analyzing the accuracy of trends, especially from a program as complex as NAEP, involves examination of many facets of the program. That is, numerous sources can contribute to variations in trends; some sources of variation are desired, while other sources contribute to noise or error. We would expect to see that "true" sources of variation are relatively larger than the variations due to noise. True sources of variation include factors such as changes in student achievement, curriculum changes, population changes, societal expectation, and student motivation arising from it. Noise can result from factors such as sampling of students, context effects, changes in test administration procedures, test assembly design, methodological changes in score reporting, and the number and nature of common items used for establishing the trend.

Since this report is part of a compendium of papers addressing many of the above and related issues, the scope of this paper is limited to investigation of trends in the subject area of mathematics. More specifically, the paper focuses on the following questions:

- Are NAEP trends in mathematics accurate?
- Are the structure and quality of exercises used for trend reporting reasonable?
- Do exercises used in NAEP instruments provide reasonable information for the variety of audiences that NAEP seeks to serve?
- What are the implications for State-by-State comparisons from the

above analysis?

The paper will examine trends by comparing NAEP trends with those available from other sources. The basic premise of the investigation is that if NAEP trends agree with trends from most of the available data, then NAEP probably provides accurate trend information. Conversely, if the NAEP trends do not agree with most other available data, then it will be difficult to derive any conclusion about the accuracy of NAEP trends.

We must acknowledge at the outset that no data set can truly be used to validate NAEP trends. One reason is that NAEP assessment is based on age, whereas others are based on grade. There are also differences in the populations assessed. NAEP is the only program that systematically reaches the sample for the defined population of test takers in the United States. Comparison of scores from various tests can also be biased by differences between tests; the skills tapped by one test might show different trends than those tapped by another. Some other differences include test administration procedures and time of testing during the year. In spite of these limitations, however, some data sources are available to compare trends with NAEP. Generally, we have richer data sources to compare NAEP trends at age 17 than at ages 9 and 13.

The data sources used for comparing NAEP's trend for the 17-year-olds include Scholastic Aptitude Test (SAT), American College Testing (ACT) program, Tests of General Educational Development (GED), National Longitudinal Study (NLS) of the High School Seniors Class of 1972, the High School and Beyond (HSB) study, and the Iowa testing programs's Iowa Tests of Educational Development (ITED). The data from the Iowa testing program's Iowa Tests of Basic Skills (ITBS) was used to judge the NAEP trends for the 9- and 13-year-olds. Throughout this examination, a liberal use of the work of Koretz (1986, 1987) has been made in providing summaries of data from various sources.

Trends in Mathematics for 17-Year-Olds

The trends reported by NAEP from 1973 through 1986 are shown in Figure 1. The trends are shown on the mathematics proficiency scale developed by ETS. The 17-year-olds showed a decreased performance between 1973 and 1982, however, they showed an upturn between 1982 and 1986. The trends for white, black, and Hispanic populations are shown in Figure 2. Black students have shown steady improvements except for a decline of scores between 1973 and 1978. Hispanic students also showed improvements, except for no change between 1973 and 1978 assessments. White students showed a continued decline through 1982, then improved significantly between 1982 and 1986.

Most of the trends reported by NAEP are supported by trends from other large-scale national data sources. For example, studies (Koretz, 1986) using NLS and HSB data report that between 1972 and 1980, mathematics achievement of high school seniors declined. As shown in Figure 3, scores on the mathematics portion of the SAT show a sharp decline between 1973 and 1978 before showing a slight upturn in 1982. From 1982 to 1986, the SAT

results have shown steady increases, as have the results from NAEP.

The 1986 Mathematics Report Card (Dossey, Mullis, Lindquist, and Chambers, 1988, p. 20), compares trends in mathematics performance for the SAT and NAEP for 17-year-olds. Both the SAT and the NAEP results show stability between 1978 and 1982, and both show modest improvement between 1982 and 1986.

Table 2 shows the trend in mathematics achievement during the period 1977 through 1985 for General Educational Development (GED) Tests. The GED trend also supports the trend reported for the NAEP. (GED Testing Service, 1988).

Koretz (1986) analyzed trends obtained from various tests according to skills in mathematics such as computation versus problem solving. His analysis showed that the average performance in mathematical knowledge did not change at all during the 5-year interval; however, understanding and applications showed declines during the same period. These results parallel those reported by the second international mathematics assessment conducted by the International Association for the Evaluation of Educational Achievement (IEA; McKnight, 1987). The IEA study showed that during the 18-year period, the declines in mathematics achievement were greater for more demanding comprehension and application items than they were for computation items at the eighth-grade level. The IEA data at the twelfth-grade level do not show the same pattern for the decline of higher-order skills as does NAEP; however, other data such as NLS, HBS, and ITED show that drops are generally in areas that are taught indirectly in schools, such as vocabulary, inferential comprehension, and problem solving.

Trends in Mathematics for Nine- and Thirteen-Year-Olds

NAEP trends for ages 9 and 13 are shown in Figure 1. At age 9, scores show an increasing trend with each successive assessment. At age 13, the scores dropped slightly between the 1973 and 1978 assessment; however, the scores have shown an increasing trend since then.

These results of NAEP are consistent with other national data sets such as ITBS. As shown in Figure 4a, Koretz (1986) reported trends for third-grade for ITBS showing a short dip accompanying an 8-year hiatus in an otherwise unbroken 30-year increase in achievement. The total decline was only 0.07 standard deviation. The NAEP results are consistent with ITBS, except that since NAEP collects data for a 4- to 5-year period, the small decline observed on ITBS was not seen on NAEP.

The results from the ITBS test also reveal that the scores of the eighth graders (Figure 4b) declined about one-third of a standard deviation at or around 1978, when NAEP's 13-year-olds showed a downward trend.

Analysis of Trend Questions

For establishing the trend across assessments, NAEP uses a small set of unreleased exercises that are common across cycles. Table 1 shows the number of common questions across three assessment years from 1977-78 to 1985-86 by scales used in the 1985-86 assessment.

A review of NAEPs mathematics framework for objectives and exercises shows that the framework has changed from one cycle to another. Perhaps these changes were made to accommodate the wishes of the mathematics committee for that particular cycle. For example, in the second mathematics assessment, the content by process matrix was defined as follows: Content--number and numeration variables and relationships; shape, size and position; measurement; other topics. Process--mathematical knowledge, mathematical skill, mathematical understanding, and mathematical application. For the fourth assessment cycle, the content by process matrix was defined as follows: Content--mathematical methods; discrete mathematics, data organization and interpretation; measurement, geometry, relations and functions; numbers and operations. Process--problem solving/reasoning, routine application, understanding/comprehension, skill, knowledge. The change in the framework for mathematics assessment shows that NAEP committee members emphasized problem solving, mathematical methods and discrete mathematics in 1985-86 assessment compared to the 1977-78 assessment.

This raises a pertinent philosophical question for the measurement of change. Can we change the framework of objectives and exercises from cycle to cycle, yet be able to measure change accurately? In designing the framework for objectives, educators in the subject matter will, rightly, reflect their concerns so that the assessment gives the proper "message" to teachers; however, such a structural change poses a threat to the measurement of trends. For future assessments, NAEP should consider balancing the two opposing, yet valid criteria. One way to resolve this dilemma is to design a much more comprehensive framework for assessment which may include all possible concerns of educators over a long period of time, say 20 years. The actual NAEP objectives in a particular cycle could be a subset of these objectives. Content changes over assessment cycles should be specified in terms of this framework.

For an accurate measurement of trends, it is also important to select a large number of common items that are stratified by content as well as by process. In reviewing the results from the second mathematics assessment, Carpenter et al. (1981) noted that "two exercises accounted for over three-fourths of the total decline in performance at age 9. These exercises involved the application of multiplication and division, which are first introduced in the third and fourth grades. These two exercises are hardly representative of the mathematics we would expect 9-year-olds to have learned, but they account for most of the change in performance" (p. 9). Perhaps these comments resulted because of the fewer exercises used for trends and the fact that two exercises showing the declines were the most difficult ones.

Quality of Mathematics Exercises

NAEP uses multiple-choice as well as open-ended questions for assessment. A review of the multiple-choice questions shows that NAEP has good quality exercises intended for the assessment of knowledge, skill, and routine application. Although there are a few good questions measuring problem-solving and understanding, it appears that NAEP can substantially improve the quality of questions measuring these processes.

NAEP's open-ended questions look like multiple-choice questions with the choices removed. NAEP may want to consider using open-ended questions that measure students' knowledge in areas that are generally difficult to measure with multiple-choice type of questions, such as students' ability to communicate, conjecture, formulate, etc. as recommended in Curriculum and Evaluation Standards for School Mathematics (1987 draft).

Conclusions

1. A comparison of NAEP's 18-year trend in mathematics established by four assessments with the trends from other data sources, such as SAT, ACT, ITED, ITBS, HSB, and NLS, shows that NAEP trends are consistent with most other trends. The magnitude of increase or decrease between NAEP assessments could not be evaluated because of the lack of information about the standard deviations.

2. NAEP trends for subgroups, such as groupings by sex and ethnicity (black, Hispanic, white), were consistent with other available data. The study established comparisons for subgroups primarily for 17-year-olds; data were lacking from other sources to make this comparison at ages 9 or 13.

3. NAEP was able to detect declining trends in scores for problem solving and thinking skills. Decline in higher-order skills, such as inferential comprehension and problem solving, was also found in other studies such as IEA and HSB.

4. Because NAEP assessments were carried out on a 4- or 5-year cycle, NAEP trends generally showed smaller increases or decreases in trend scores as compared to other data sources.

5. For mathematics, the nature and quality of NAEP questions is similar to those on standardized tests, such as ITED, ITBS, and CAT. Compared to questions on IEA, NAEP questions were simpler, more traditional, and lacking in items to assess understanding and problem solving.

6. An examination of common items for the trend analysis shows that perhaps the selected items are stratified by content domain but are not stratified by process/difficulty.

Recommendations

1. Common items for trends. Common items must at least be stratified on content and difficulty. There will be other statistical criteria for selecting the common items for trend. The overall trend can be significantly biased by choosing all easier or all difficult items or by choosing mostly items that assess computation and understanding versus mostly problem solving.

2. Domain definition. In the past few years, NAEP has reported student achievement in mathematics skills such as in computation, understanding, and problem solving. There is a need to think of a new reporting taxonomy that could be the basis for collecting and reporting mathematics scores.

The achievement in any one subject can be defined and measured in many different ways, and the variations in measurements can be large enough to create very different trends. The reporting unit should be able to tap differential trends in student learning embodied in any reform effort in general and recent curricular reforms in particular. The categories of reporting should not only be based on traditional content by process matrix but also be based on knowledge acquisition theories.

3. Quality of exercises. The variety and quality of exercises should be improved significantly. NAEP exercises should include multiple-choice, open-ended, and performance type of questions as recommended in NCTM's Curriculum and Evaluation Standards for School Mathematics.

4. Comprehensiveness. If NAEP exercises will also be used in instruments for State-by-State comparisons, the NAEP instruments will have the status of a single national achievement test. The comparison of various tests and trends derived from them has shown that a variety of measures are often needed to reach reasonable conclusions about student achievement. There is great danger of being misinformed by a single test, because it is often impossible to foresee when a single test will be misleading.

Furthermore, if the NAEP test ever achieves the status of national test, NAEP exercises will be subjected to closer scrutiny by curriculum coordinators, and instruction in schools is likely to be tailored specifically to raise scores. Therefore, the exercises must not only have good psychometric properties, but also serve as exemplars of good teaching practices. The test content should be sufficiently comprehensive and balanced; the test should neither be narrow in content nor should it distort the curriculum.

References

Beaton, A. E. et al. (1987). Implementing the new design: The NAEP 1983-84 technical report. (NAEP Report 15-R-01). Princeton: Educational Testing Service.

Carpenter, T. P. et al. (1981). Results from the second mathematics assessment of the National Assessment of Educational Progress. Reston, VA: National Council of Teachers of Mathematics.

College Entrance Examination Board. National college bound seniors. New York: The College Board.

Dossey, J. A., I. V. S. Mullis, M. M. Lindquist, D. L. Chambers (1987). The mathematics report card: Are we measuring up? Trends and achievement based on the 1986 National Assessment. Princeton: Educational Testing Service (draft).

General Educational Development Testing Service (1988). Unpublished data on trends in performance of graduating high school seniors on the Tests of General Education Development. Washington, DC: American Council on Education.

Iowa Testing Programs (1985). Iowa Basic Skills Testing Program, Achievement trends in Iowa: 1955-1985. Unpublished.

Koretz, D. (1987). Educational achievement: Explanations and implications of recent trends. Washington, DC: Congressional Budget Office.

Trends in educational achievement (1986). Washington, DC: Congressional Budget Office, 1986.

McKnight, C. C. et al. (1987). The underachieving curriculum: Assessing U.S. school mathematics from an international perspective. Champaign, IL: Stipes Publishing Co.

National Assessment of Educational Progress (1985). Mathematics objectives, 1985-86 assessment. Princeton: Educational Testing Service.

Trends in mathematical achievement, 1975-78 (1979). Denver, CO: Education Commission of the States.

National Council of Teachers of Mathematics (1987 draft). Curriculum and evaluation standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics.

Table 1

Number of Common Questions Across Three Assessments
by Scales in Mathematics

	No. of questions	1977-78			1981-82			1985-86		
		9-	13-	17-	9-	13-	17-	9-	13-	17-
Numbers & Operations -- Knowledge	3	x			x			x		
	11		x			x			x	
	5		x	x		x	x		x	x
	10			x			x			x
	6				x			x		
	1					x			x	
	2				x	x		x	x	
	3	x	x		x	x		x	x	
	1						x			x
Numbers & Operations -- Applications	5				x	x		x	x	
	6			x			x			x
	4		x			x			x	
	3		x	x		x	x		x	x
	3	x			x			x		
	5					x	x		x	x
	2					x			x	
	1				x			x		
Fundamental Methods	1	x	x		x	x		x	x	
	2			x			x			x
	3	x	x	x	x	x	x	x	x	x
Relations & Functions	7			x			x			x
	3		x	x		x	x		x	x
	2	x			x			x		
Geometry	1						x			x
	2		x			x			x	
	4			x			x			x
	5		x	x		x	x		x	x
	1	x			x			x		
Measurement	5			x			x			x
	4				x			x		
	2				x	x		x	x	
	6					x			x	
	1					x	x		x	x
	2	x	x		x	x		x	x	
	1		x	x		x	x		x	x
	5	x			x			x		
	1	x	x	x	x	x	x	x	x	x
Data Organization	3				x			x		
	8	x	x		x	x		x	x	
	5			x			x			x
	1		x	x		x	x		x	x
	1		x			x			x	

Table 2

Performance of Graduating Seniors on Anchor Form (MA) of the Tests of General Educational Development (GED) by Year: Raw Score Means (Standard Errors) and Standard Deviations

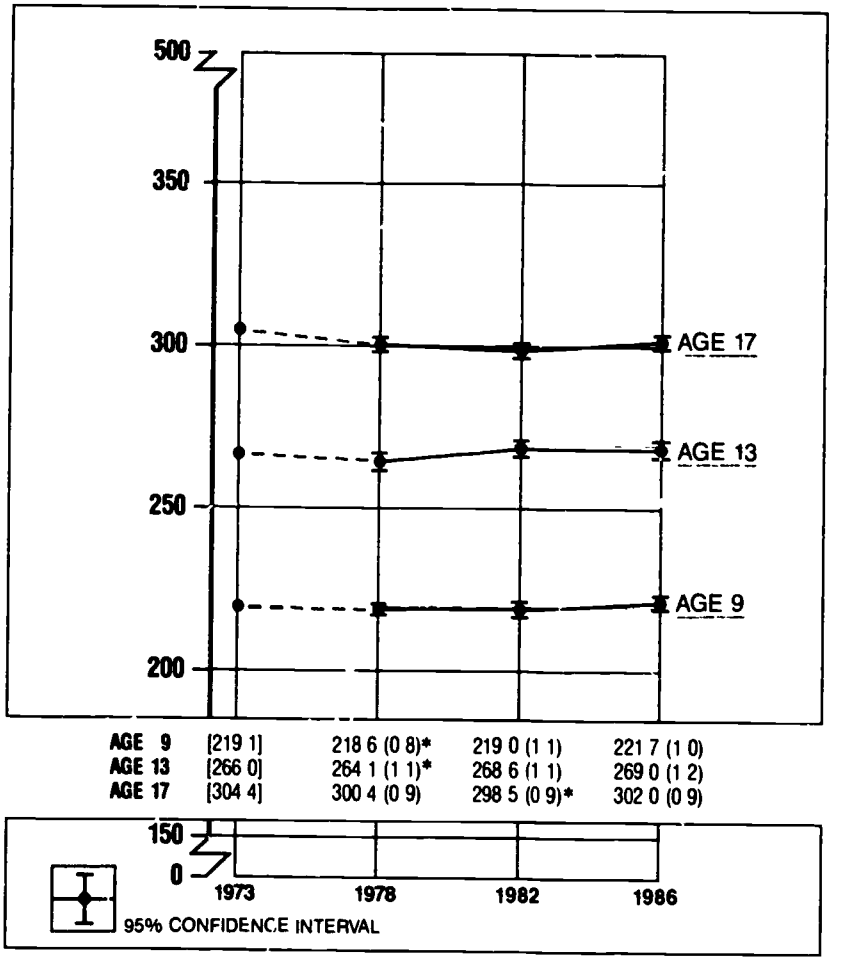
Test (# of Items)	Statistic	Year			
		1977	1980	1983	1985
Writing (80)	Mean (SE)	48.9 (.23)	47.1 (.25)	44.9 (.61)	47.4 (.58)
	SD	14.0	14.9	16.2	15.4
Social Studies (60)	Mean (SE)	38.2 (.20)	36.7 (.20)	35.2 (.44)	35.6 (.44)
	SD	11.7	12.0	11.5	11.6
Science (60)	Mean (SE)	34.5 (.19)	33.3 (.19)	33.9 (.46)	33.3 (.42)
	SD	11.2	11.4	11.8	11.0
Reading Skills (40)	Mean (SE)	27.4 (.14)	26.6 (.14)	27.2 (.30)	26.5 (.32)
	SD	8.2	8.3	7.7	8.3
Mathematics (50)	Mean (SE)	30.1 (.15)	29.1 (.16)	30.4 (.36)	30.4 (.35)
	SD	9.2	9.4	9.3	9.1

Note: 1977 and 1980 results are from GED norming studies conducted in those years. The 1983 and 1985 results are from GED equating studies.

Figure 1

National Trends in Average Mathematics Proficiency for 9-, 13-, and 17-Year-Olds: 1973-1986

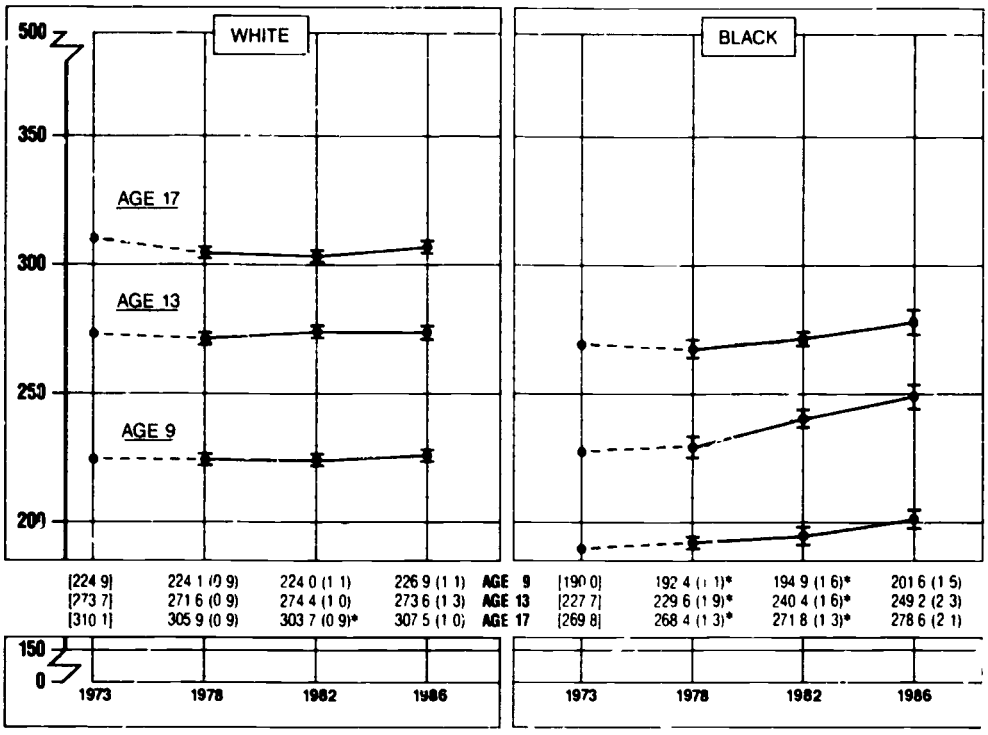
FIGURE 1 1

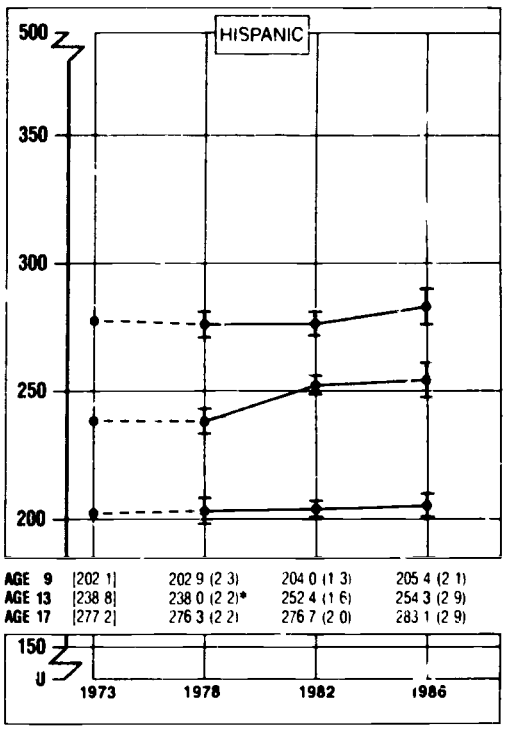


(Reprinted from The Mathematics Report Card: Are We Measuring Up?, 1988, p. 19)

Figure 2

Trends in Average Mathematics Proficiency for 9-, 13-, and 17-Year-Olds by Race/Ethnicity: 1973-1986





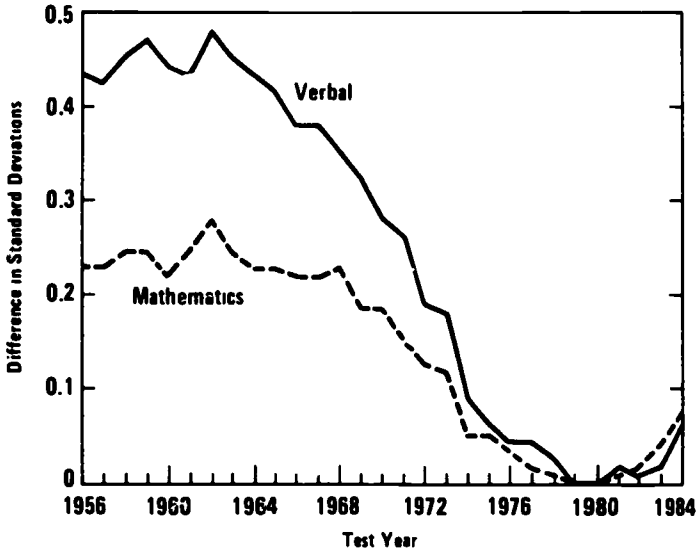
[– – –] Extrapolated from previous NAEP analyses
* Statistically significant difference from 1986 at the .05 level
Jackknifed standard errors are presented in parentheses

95% CONFIDENCE INTERVAL

THE NATION'S REPORT CARD

Figure 3

Figure III-4.
Average SAT Scores, by Subject, Differences from Lowest Year

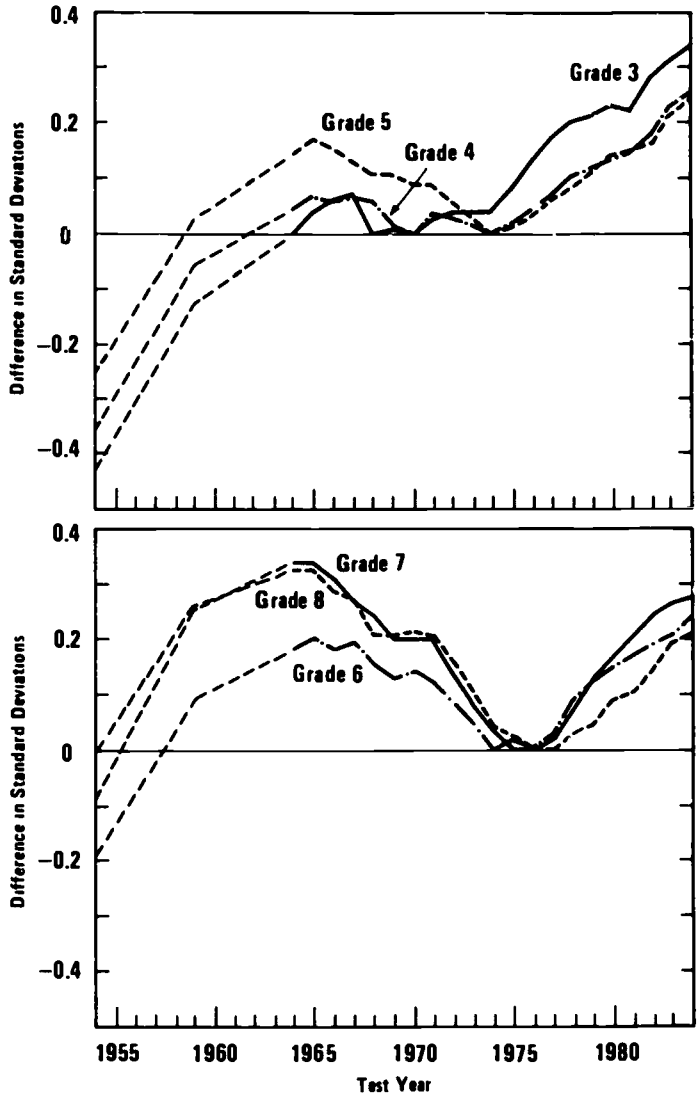


SOURCES: CBO calculations based on Hunter M. Breland, *The SAT Score Decline: A Summary of Related Research* (New York: The College Board, 1976), Table 1, and the College Entrance Examination Board, *National College-Bound Seniors, 1985* (New York: The College Board, 1985).

(Reprinted from Koretz, 1986, p. 38.)

Figure 4

Iowa Composite, ITBS, Grades 3-8, Differences from Post-1964 Low Point



SOURCES: CBO calculations based on "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985" (Iowa Testing Programs, unpublished and undated material), and A. N. Hieronymus, E. F. Lindquist, and H D Hoover, *Iowa Tests of Basic Skills: Manual For School Administrators* (Chicago: Riverside, 1982).

(Reprinted from Koretz, 1986, p.63.)

Quality Control: The Custodian of Continuity in NAEP Trends

William H. Schmidt
National Science Foundation

The establishment and interpretability of data trends depend not only on statistical and sampling consistency, but on procedural consistency as well. NAEP has been subjected somewhat routinely to the introduction of new procedures, usually for what the purveyors believe to be very good reasons, but with little consideration for the impact such changes can have on the comparability of the data over time.

The greatest priority for a testing procedure that is designed to be the nation's report card is consistency. Imagine the outcry among parents if schools changed the process underlying report cards every semester. Administrative procedures should only be changed when there is a compelling reason to do so and then only after extensive deliberation to determine if the reasons make worthwhile the risks to continuity attendant to procedural changes. This does not imply that such changes should never be made, only that they be done after an extensive examination of the likely consequences.

Proposed changes which survive the rigorous deliberative process outlined above, must then be subjected to careful empirical examination to determine their likely effect on the testing procedure. Such bridging studies should be done prior to and not concomitant with the introduction of the changes into the main data collection. This permits the a priori estimation of the magnitude and direction of the effects of such a change on the trend data; the results of which can then be considered in the deliberative process weighing these distortions to continuity against the relative advantages of the new procedures.

Past Experience

The reading anomaly is only the latest in a series of awkward situations for NAEP that suggest ill-considered methodological changes or faulty design decisions. In fact, not only are the current reading scores questionable but the trend lines for scores in other subject matters could also be called into question. It is perhaps fortunate that the anomaly in reading between 1984 and 1986 was large enough to detect since other less obvious results might not have drawn our attention to the potential problem. In fact data points in other years and for other subject matters might be flawed by a lack of procedural consistency.

An example of one ill-considered methodological change involves BIB spiralling. It is clear the BIB spiralling was overdone in 1984 when ETS discovered after the fact that they could not get estimates of ability for a large proportion of examinees because so few items in a given content

area had been answered by each examinee. The result was a retreat to "plausible values," which are based on a student's demographic group as well as exercise responses. After "plausible values" were implemented, it was discovered that serious biases arise if you attempt to use them to get breakdowns of achievement according to demographic categories not used in the original conditioning. It would appear ETS rushed into BIB spiralling, probably went too far with it and as a result was forced into a series of costly and not entirely satisfactory methodological accommodations.

The Reading Anomaly of 1986 - The Breakdown on NAEP Trends

The reading anomaly in 1986 appears to have arisen from a large number of untested modifications in test administration procedures. The purpose of this section is not to determine the relative contribution of each of these modifications, but to illustrate how a series of well-intentioned but ill-conceived methodological changes or faulty design decisions could have led to the reading anomaly (a more through analysis of the factors can be found in the paper by Hedges).

For example, in 1984, NAEP booklets containing reading items consisted of one to three blocks assessing reading and one to three blocks assessing writing to yield a total of three blocks per booklet. In 1986, assessment booklets were also divided into three blocks. The non-reading blocks included items in the content areas of mathematics, science, computer competence, history, and literature. The following is taken directly from Beaton et al.:

> Students at all three ages in the 1984 BIB-spiralled assessment sessions took booklets containing three blocks. These booklets contained 0, 1, 2, or 3 reading blocks; the remaining blocks, if any, consisted of writing exercises. The students had to read some instructions and the exercise texts. The 1986 bridge assessment for ages 9 and 13 contained three booklets, each of which contained three blocks. The subjects were math, science, and reading and the booklets were configured as shown in Table 7-2.
>
> The same booklet was administered to an entire assessment session. The math and science parts of the booklets were paced (presented aurally using a tape recorder). The tape recorder was turned off for the reading block in each session. For age 17, the BIB booklets in 1986 contained 0, 1, 2, or 3 reading blocks; the remaining blocks, if any, were in math, science, computer competence, or, in the case of 4 of the 97 booklets, history and literature. In 1986, the age 13 and age 17 reading blocks were identical in every respect so that the three blocks (13R1, 13R2, and 13R3) used in the age 13 bridge were repeated as part of the age 17 BIB reading blocks. Different students in the same session were administered different booklets.
>
> The length of time allotted for each block changed between 1984 and 1986. In 1984 each age was given a six-minute common

core of background and attitude questions followed by three subject area blocks of fourteen minutes each. At the end of each fourteen-minute interval, the students were told to move to the next block. Approximately the first two minutes of these subject area blocks were devoted to answering additional attitude questions related to the curriculum area. In 1986, the age 13 and 17 students again had six minutes to respond to the common core background and attitude questions; however, for 9-year-olds, the common core questions at the beginning of each were read aloud to them and took 15 minutes to complete. The 9-year-olds were given 13 minutes to read and respond to the exercises in the block; the 13- and 17-year-olds were given 16 minutes.

Hedges (this report) estimates that these changes alone could account for nearly half the size of the anomaly.

In addition to these changes in booklet format and administration, the report by Beaton et al. considered six other general classes of potential explanations for the anomalous score decline. Included are changes in: the population tested, scoring procedures, and administrative procedures such as the date of the testing, and the number of individuals assessed in each session. Other classes of explanations for the score decline include: quality control, shifting item non-response patterns and artifacts of scaling. Since the focus in this paper is on procedural continuity, we discuss only changes in the population tested, scoring procedures, and administrative procedures.

Consider first the population tested. Good intentions were behind the change in the NAEP sampling design. As the Beaton et al. report suggests the design was "improved in a number of ways" to increase the "power of NAEP data as well as to increase statistical and administrative efficiency." Still design changes were introduced into the 1986 NAEP threatening at least in principle the desired continuity. Post-hoc studies suggest that these changes did not likely contribute to the anomaly. This however, was learned only after-the-fact.

Administrative changes were also introduced into the 1986 NAEP assessment. The average number of individuals assessed in each session was increased from 20 to 35 for the 17-year-olds. Also during the assessment session for 17-year-olds, up to 5 teachers were introduced. The dates of the data collection activity were also changed between 1984 and 1986.

The 9- and 17-year-olds were tested earlier in 1986 than in 1984, while the 13-year-old sample was tested later. The average difference in dates between 1986 and 1984 corresponds to -22 days for 9-year-olds, +4 days for 13-year-olds, and -18 days for 17-year-olds.

Changes in scoring procedures were introduced between 1984 and 1986. In the 1986 assessment booklets students were asked to "fill in the oval" and responses were then machine scored. By contrast the 1984 assessment booklets instructed students to "circle the letter" and responses were key

entered.

All or some of these changes likely combined to produce the reading anomaly. The procedures introduced in 1986 seem not to be unreasonable on the surface and represent by themselves acceptable test design or administration procedures for the most part. But the problem is that they represent changes from NAEP's previous-years procedures; the introduction of such methodological changes into the main data collection without first testing their likely impacts, and them considering them in light of the major priorities of NAEP is what is ill-considered.

Quality Control

Quality Control we define as all that must be done to insure the continuity of the process and as a result the NAEP trends. Such efforts must not only be in support of administrative continuity as outlined in the foregoing paragraphs, but must also extend to a continuity in the framework used to define the content domain. The domain should remain stable and the knowledge and skills assessed each time should be drawn from that domain. It appears from past assessments that NAEP objectives have been developed ad hoc for each successive assessment with little attention to consistency over time.

Quality control needs to be taken much more seriously. It needs to include a systematic procedure for considering all changes from one assessment cycle to the next in exercises, in the composition of exercises within a booklet, in administration conditions and in scoring and data coding formats. The possibility of such changes need to be considered in light of NAEP priorities. This implies that changes should be made only when doing so is critical to achieve the priorities; to change only because a new procedure is technically preferable is not sufficient. In fact, since change is a threat to the continuity of the NAEP time series, itself a high priority, the change should be considered only if the priority it serves is as high or higher than continuity.

What Should Be Done?

NAEP needs a systematic quality control mechanism; not only for its exercises, but the contexts in which they are administered need to be controlled. Design innovations need to be piloted before being used on a large scale, and the entire analysis plan needs to be set forth and thoroughly critiqued before data collection begins. NAEP needs to bring to bear all relevant areas of expertise--the current ETS replications of the 1984 and 1986 procedures are a vital source of information for future design decisions. However, such ad hoc investigations are not sufficient to assure continuity of NAEP trends. A new process should be developed to ensure adequate and systematic evaluation of proposed procedural changes. The technical advisory process to NAEP should comprehensively incorporate considerations of procedural design and audit as well as sample design and analysis. This implies formal review of on-site administration conditions and procedures, instructions and student conformity to them, etc., as well as timing and booklet design. This also implies that the technical

advisory body should be composed of individuals representing all relevant areas of expertise.

Undoubtedly, there will be profound changes in future NAEP data collections, especially in light of recommendations to extend NAEP to permit State-by-State comparisons. However, whatever modifications are made in the overall design it is mandatory that the procedures used to collect the data for national trend estimates be parallel in every important respect. During transitions when old and new procedures are carried in parallel, not only the assessment exercises themselves but also the data collection procedures should remain the same.

Quality control would seem to be important to both developmental activities including exercise development, response and booklet formats, and studies of different data collection procedures; and validation activities including statistical studies of error sources and design control. The Alexander-James report, The Nation's Report Card, called for establishing an independent Educational Assessment Council (EAC). Included in their responsibilities is the selection of the content areas to be tested, and setting policy on "such matters as maintaining the continuity over time of the assessments data banks." Within this purview would certainly fall the responsibility of setting policy for quality control and being the body to systematically examine the likely consequences of any proposed changes in administrative procedures and considering them in light of NAEP's priorities.

Assessment of National Trends in Achievement: An Examination of Recent Changes in NAEP Estimates

David E. Wiley
Northwestern University

1. Introduction

The anomalous decline in National Assessment of Educational Progress (NAEP) reading test scores for 9- and 17-year-olds between 1984 and 1986 is not a uniform decline in level for these two age groups. These drops in level took place in the context of an increase in performance variability which also occurred for 13-year-olds. Changes in variability can occur for substantive reasons--e.g., reallocation of instructional resources from pupils doing poorly to those who are doing well--or for methodological reasons--e.g., less standardized testing conditions. Thus, it is important to account for these changes in order to understand why the results of the assessment may be different from one cycle to the next.

Another aspect of the decline is that it has not been paralleled by similar changes in Scholastic Aptitude (SAT) or American College (ACT) test scores. In fact, however, one would not expect parallelism with these other indicators. The NAEP 17-year-olds are approximately the same age as the college-bound seniors who take the SAT or ACT, but one would expect trends to differ for two primary reasons. (a) The groups tested in the college entrance examination programs are subgroups of the population sampled by NAEP. And these subgroups are self-selected on the basis of ability and high school performance. Clearly, the national trends for higher performing students need not parallel those of average performers. (b) Secondly, the content of the testing instruments used is not the same. For example, the SAT verbal test has a vocabulary subtest, which is not true of NAEP. Both the SAT and NAEP use a balanced selection of reading passages of various types but these types differ in important ways. Also, the typical difficulty levels of the passages in the two tests differ significantly.

In this paper, I attempt to evaluate these trend differences and to assess the implications for interpretation of the anomaly and for the future conduct of NAEP.

2. Specifying the Anomaly: Change in Distribution vs. Change in Level

Beaton (1988), in the ETS report on the reading anomaly, reports a drop of 6.0 scale score points for 9-year-olds, a gain of 2.4 points for 13-year-olds, and a loss of 10.7 points for 17-year-olds. These are accompanied by increases in the standard deviations of the score distributions of 10, 11, and 25 percent, respectively. These increases in variability signal that the mean level changes do not all tell the whole story about changes in performance.

In order to examine these changes in more detail, I computed selected percentiles of the distribution for both years for each age group, using score distribution data computed by ETS for their report and released to the panel for further study. Below, in Table 1, I display the differences between the scale score values for the two years at various percentile levels. These values indicate the difference in achievement, in scale score points, of pupils in 1986 and 1984 who were at the same relative percentile in the 1986 and 1984 distributions, respectively. Thus, for example, at the 95th percentile, 17-year-old pupils in 1986 had scores 9.5 points higher than similar pupils in 1984. At the median, however, the scores are 8.65 points lower, which corresponds to the reported mean drop of 10.7 points.

Table 1.--NAEP Reading Scale Scores--Differences in 1986 and 1984 Percentile Values

	Values		
Percentile	9 Years	13 years	17 years
99	2.67	11.01	14.67
97	1.51	8.45	9.51
95	2.43	8.00	9.47
90	0.39	8.39	6.45
80	0.67	7.31	2.14
75	-0.96	8.84	0.60
50	-2.57	4.89	-8.65
25	-6.19	2.00	-17.53
20	-6.56	0.16	-10.44
10	-7.74	1.50	-24.33
5	-17.68	-0.28	-24.45
3	-11.30	0.00	-28.00
1	-8.53	-4.00	-26.27

In general, at sufficiently low percentiles there are score declines for all age groups. Also, at high percentiles, there are gains for all age groups. Thus the general finding, in terms of reported scores, is that high ability children are performing better, and low ability children are doing worse in all age groups. The major difference among the age groups is the "stable" percentile point, i.e., that point below which losses occur and above which gains occur. This point is at about the 80th, 5th, and 75th percentiles, for 9-year-olds, 13-year-olds, and 17-year-olds, respectively.

In general, the potential causes of such distributional changes, fall into three categories: (a) methodological artifacts, (b) changes in population, and (c) changes in pupil learning. Methodological artifacts include changes in testing procedures, changes in the sampling frame, or in the implementation of the sample. It appears most likely to me that because substantial changes in the packaging of exercises occurred between 1984 and 1986--causing changes in response context and in subset timing--

that any methodological artifacts most likely derive from these changes.

The second possibility is that the populations from which the sampled age groups were drawn changed over the 2-year period. Over longer periods of time, such changes have occurred in the past. Differential birth rates in subpopulations and immigration can cause significant demographic shifts in population, changing the family backgrounds of age groups of pupils. However, these shifts are slow and a two year period is not likely to exhibit much change of this kind.

The last possibility is actual changes in learning--either in or out of school. Again, however, changes in instruction or access to learning environments usually would not result in substantial national achievement differences over a 2-year period unless simultaneous programmatic changes were made in many schools located across the Nation. Also, the distributional pattern would indicate that such schooling changes are producing higher scores at the top of the distribution and lower scores at the bottom. A shift in instructional resources from low achievers to high achievers would have this result. Many observers have seen signs consistent with this possibility in that programs for the gifted have been advancing while resources allocated to the educationally disadvantaged have gradually drifted downward. However, the 2-year time period is still quite short for an achievement effect of the magnitude reported. In my opinion, the magnitude of these declines over this short time period together with the accompanying large increases in variability point to the conclusion that the most likely cause is methodological.

3. Population Differences between NAEP and Other Test Programs

The SAT and the ACT are the only testing programs which report time trends in performance besides NAEP. However, the college entrance testing programs serve only those who aspire to attend 4-year colleges or universities. These individuals still constitute a minority of high school seniors and generally are those who have better secondary school performance than non-entrance test takers. Consequently, the likelihood of a high school senior taking, e.g., the SAT is greater if the senior's ability is high rather than low. Generally, then, we would expect participation in a college entrance testing program to be an increasing function of ability.

If the likelihood of test participation is close to certain for students with high enough ability, the upper end of the ability distribution of NAEP participants will be similar to the upper ends of the college entrance test program participants. Following this reasoning to its logical conclusion, one can formulate procedures to compare achievement trends of high ability students using the overall distributions of NAEP performance scale scores and college entrance scores. I have attempted this for NAEP and the SAT.

For the SAT, I am willing to assume that some regularity exists at the highest achievement levels. If in the selected population the selection ratio at high abilities is close to 100 percent, then an estimate

of the top percentile achievement levels in the unselected population can be estimated. For example, if all of those above the 90th percentile in the total population were selected into the SAT group, and if 40 percent of the total population took the SAT, then the 90th percentile in the overall population would correspond to the 75th percentile in the SAT group [(1 - .9)/.4 = .25]. In general, assuming 100 percent selection above a specified level, the percentile, $\pi(\alpha)$, in the unselected population, corresponding to achievement level α, will equal $1 - p(i)[1 - \pi^*(\alpha)]$, where $\pi^*(\alpha)$ is the percentile in the selected population and $p(i)$ is the overall proportion selected. This follows, under the assumption of 100 percent selection above $\pi(\alpha)$, since then $[1 - \pi(\alpha)/p(i)] = 1 - \pi^*(\alpha)$. Thus, one may adjust the SAT over years to constant percentiles in the overall distribution and compare performance levels with NAEP at those points. E.g., the 99th, 97th, 95th, and 90th.

In terms of the anomalous decline in 1984-86 reading scores, the maximally appropriate comparison with SAT trends is in the Reading Comprehension subscale of the verbal scale score. The other verbal subscale, vocabulary, has no direct correspondent in NAEP.

In applying corrections to these SAT distribution data, I have used figures which estimate the number of twelfth graders from the Digest of Education Statistics, 1987 as well as SAT distribution data taken from the College Board publications: National Report, College-Bound Seniors, 1980, 1984 and 1986. The ratio of the number of college-bound seniors taking the SAT to the total number of twelfth graders was used to calculate the percent of the NAEP population which took the SAT in each year. [Note: the number of twelfth graders in 1986 was estimated by applying the 12/11 ratio in 1985/1984 to the 1985 eleventh grade enrollment value.] The resulting percentages of seniors taking the SAT were 33.9, 37.1 and 38.6 in 1980, 1984, and 1986, respectively. Correspondences were computed by selecting percentile values which produced approximate equivalences for the latter two years since there was a relatively small change in corresponding percentiles from 1980 to 1984 and almost no change from 1984 to 1986.

Table 2 displays the percentile equivalences calculated from these percentages together with the corresponding SAT scale score values for 1980, 1984, and 1986. It should be noted that the average SAT values for college-bound seniors increased over this period. Also, the SAT values corresponding to the NAEP 99th, 97th, 95th, and 90th percentiles increased, albeit irregularly. Since the estimated NAEP values for these percentiles also increase (Table 1), there is no obvious inconsistency between the SAT trends and the NAEP trends.

4. Content Analysis Comparisons of SAT and NAEP

A content comparison between Scholastic Aptitude Test and the National Assessment of Educational Progress is not direct. One reason for this is that NAEP has specified its content in terms of "objectives," while the SAT phrases its content in terms of item specifications. In theory, the distinction between objectives phrased in terms of intended abilities and item categories based on the properties of stimulus materials and questions

is great. However, in practice, the categories actually used tend to be eclectic and are labeled ambiguously so that both ability and test task interpretations are mixed.

Table 2.--Reading Comprehension--Approximate Percentile Equivalences Between NAEP and SAT with Corresponding SAT Scale Scores

Percentile		SAT Scale Score Values		
NAEP	SAT	1980	1984	1986
--	0.99	699	713	716
0.99	0.97	662	666	670
--	0.95	632	635	641
0.97	0.92	596	598	609
--	0.90	583	587	593
0.95	0.87	562	569	576
0.90	0.73	497	502	509
?	0.50	425	431	435
Mean SAT		425	428	433

Judged in terms of reading-related content, the SAT verbal scale has two components: vocabulary and reading comprehension. The NAEP reading scale, however, only includes comprehension questions. At the minimum, this means that NAEP reading scale score trends should be compared only to trends in the reading comprehension subscale of the SAT.

At a more specific level, reading comprehension questions traditionally have two separable parts: (a) a passage containing textual material containing verbal content to be comprehended, and (b) questions--usually multiple choice--about the text. Both the NAEP and the SAT comprehension questions are of this type. As a consequence, the differences between NAEP and SAT can be described in these terms as well.

Generally, the textual materials from NAEP are somewhat broader in scope than those in the SAT. Both tests contain material which is literary, cultural, scientific and social in nature. However, NAEP materials also include additional content such as advertisements and forms which are not used for the SAT. In addition, the textual material on the SAT is linguistically more complex, both at the sentence and paragraph levels.

The consequence of these differences in content for the abilities measured is likely to be important for average high school students. My guess is that the items of low to average difficulty in the NAEP have little correspondence to SAT items, while the more difficult NAEP items have a closer resemblance. If on the other hand, some of the non-standard NAEP items are also difficult, there may be important differences in the abilities measured by the two scales even for high-ability students.

The other content facet concerns the questions which are asked about the passages. Here the question types seem very similar. NAEP discusses analysis, interpretation, and evaluation. NAEP also refers to the location of specific items of information, making inferences, and recognizing the main idea. Similar phrases are used in the SAT content specifications.

The main issues of content comparisons are beyond the scope of a short-term panel. In order to adequately compare the content of a NAEP assessment and that of the ACT or SAT, three things must be done:

(a) a common content framework must be formulated which includes a corresponding set of content categories from both testing programs.

(b) a representative set of items from each source must be categorized and the differences in item distribution analyzed.

(c) a stratification of the NAEP population should be accomplished which would allow the selection of a group of high school seniors similar to the college entrance examination groups. Item difficulties could then be tabulated by content categories for this subgroup.

Such an analysis would indicate whether the NAEP and SAT scales are approximately equivalent <u>for the kinds of students who take college entrance examinations</u>. My tentative conclusion is that these scales are not equivalent for typical students but may be approximately equivalent for those of higher ability.

5. Conclusions

The primary conclusions of my investigation are:

(a) the observed drops in reading comprehension of 9- and 17-year-olds between 1984 and 1986 are not uniform. In fact, at the highest ability levels, students in all three age groups had higher scores in 1986 than in 1984, while the lowest ability students in each group had lower scores.

(b) after grossly accounting for selective character of the SAT test takers, corresponding groups of NAEP participants showed similar trends (gains) between 1984 and 1986.

(c) test content of moderate difficulty--for typical high school students--is sufficiently discrepant between NAEP reading and the SAT verbal scales, so that direct trend comparisons are treacherous. However, for high ability students, the reading comprehension subscale of the SAT <u>might</u> be sufficiently comparable to the NAEP reading scale to make trend comparisons of these students meaningful.

(d) the magnitude of the NAEP reading scale score changes between 1984 and 1986 together with the large increase in variability make methodological changes between the two assessments the most likely primary cause of the decline.

Given these conclusions, I enthusiastically endorse the four recommendations made by the panel on February 24-25, 1988. Similarly, I strongly endorse the two resolutions drafted and voted on by the first subcommittee on February 25. [Note: These four recommendations now appear as Recommendations 1, 5, 8 and 9. The two resolutions have been revised, and now appear as Conclusions 1 and 2.]

Management and Administration of a State-NAEP Program

Mark D. Musick
Southern Regional Education Board

Establishing and administering a nationwide student testing program that uses the National Assessment of Educational Progress to provide information on a State-by-State basis is a manageable task. The real barriers to establishing a program to provide this kind of student achievement information have been primarily philosophical or political. We have not had comparable State-by-State information on student achievement because educational leaders did not want it, or to be kinder, did not place any value on having it, and because political leaders did not demand it. In an era of educational reform and improvement in the Nation, both of these factors have changed. Educational leaders are at least agreeable to having comparable student achievement information among the States. And government leaders at the State level are now very much interested in it. Discussing the reasons for this dramatic change in attitudes is not the purpose of this paper. But without those attitude changes, the subject of management and administration of a nationwide program to provide comparable State-by-State student achievement information would be merely an academic question.

The underlying premise of a new State-based National Assessment of Educational Progress program--hereinafter referred to as State-NAEP--is that, except for the scale of operation, a national assessment of educational outcomes reporting at the State level is not fundamentally different from the present National Assessment reporting at the national and regional level. However, it is not a simple step to move from this premise to a full-scale operational testing program that would assess several hundred thousand students from perhaps most of the States in the country. A successful, fully operational State-NAEP program will involve more States and several times more students than the current 90,000 students involved in the National-NAEP program.

The State-NAEP program should be a program unit in the National-NAEP arrangement. State-NAEP might be viewed by some as a subsidiary of National-NAEP, but in this case the scale of operation of the subsidiary may soon dwarf the parent company. The scale of operation notwithstanding, National-NAEP should be seen as providing the leadership on which the State-NAEP program should be based.

The State-NAEP program will need a staff director and staff within the National-NAEP arrangement. Because of the scale of the operation, State-NAEP must have staff whose sole responsibility is to the State-NAEP program. State-NAEP will also need an advocate within the National-NAEP arrangement. That advocate should be the staff director and the advisory structure for State-NAEP.

The governance and advisory structure for National-NAEP should reflect State-NAEP interests and provide for special ad hoc advisory groups when needed by State-NAEP. There will be overlapping interests in National-NAEP and State-NAEP programs, but the purposes of the two programs are substantially different. The advisory structure for National-NAEP as recommended by the Alexander-James Report provides for a significant degree of State-level input. This may reflect in part the new charge that the Alexander-James group recommends for National-NAEP and it may reflect that some State officials believe that the current opportunities for State input into National-NAEP are unsatisfactory.

It is important that State-NAEP be seen as a program that is conducted "in conjunction" with National-NAEP. If conducted properly and "in conjunction" with National-NAEP, the State-NAEP program can gain an important head start in terms of public credibility and acceptance. Being conducted "in conjunction" with the National-NAEP means that the State-NAEP procedures will be compatible and that the assessments will be in the same year and in the same testing time frame. Conducting State-NAEP and National-NAEP in the same year has two primary benefits. First, the so-called "testing burden" is probably reduced if State-NAEP and National-NAEP occur jointly in alternate-year testing cycles. This would mean a NAEP testing program would be in the field in most States every other year. Such an arrangement would concentrate the testing schedule and provide for fewer administrative problems and classroom disruptions. As both State-NAEP and National-NAEP will test relatively small samples of students in each State, the programs will not add significantly to the "testing burden." Testing a relatively small sample of students (a sample of only a few thousand students per State at three grade levels) is a reasonable commitment to ask States to make to obtain never-before-available information on student achievement. The second reason to test in the same year with both State-NAEP and National-NAEP is that this will provide States with the most current national information on student achievement. States will be able to compare their students' performance in a given year with the performance of students nationally, and in other States, for that same year. Currently, when States compare their students' achievement to that of so-called national averages, they are often comparing themselves to national averages that are several years old. This is one of the reasons that nearly all States are currently able to show that their students' achievement is above the national average at a given grade level.

In this national election year the public is given almost weekly updates on political polls for State and national offices. The real value of this updated information about political campaigns, at least the value to the public, is probably questionable. To think that a nation would have this kind of current information about political campaigns and would settle for outdated information about educational achievement of its youngsters would seem to be an untenable position. The State-based National Assessment program as it is proposed would clearly provide the most up-to-date information available at the State level on student performance.

In order for the State-NAEP assessments to be compatible with the National-NAEP, it will be necessary that the State-NAEP testing instruments be replicas of the National-NAEP. Ensuring that State-NAEP testing instruments are compatible with National-NAEP should not be a difficult task. If the testing instruments are not exactly the same, however, care must be taken to see that the State-NAEP tests are in fact compatible with National-NAEP. The problems with the 1986 reading results from National-NAEP show the importance of attention to detail in the testing instruments and procedures.

It follows that State-NAEP tests should be constructed under the direction of the National-NAEP program. The emphasis in this paper is on the form of the tests themselves rather than on the content. Suffice it to say that National-NAEP should continue to be based on the judgment of a broad cross-section of American citizens about what students should know. National-NAEP and therefore the State-NAEP should not become a minimum competency test, nor a test requiring agreement by representatives of all States. Clearly "what is tested" must remain a major part of the focus of the National-NAEP program. By recommending that State-NAEP tests be constructed under the direction of the National-NAEP program, I am suggesting that once the content areas have been agreed upon then the test instruments, i.e. the booklets of test questions, be the responsibility of the National-NAEP program. Issues of compatibility and credibility are both at stake in decisions about how the testing instruments are constructed and who is in charge of the process.

Another compatibility and credibility issue is the preparation of the sample of students to be tested in each State. State samples should be drawn in accordance with National-NAEP procedures under the direction of the National-NAEP program. That does not necessarily mean, however, that there is not a role for States to play in assisting in this process. Preparing the sample of students from each State requires listings of schools and rosters of students. It is reasonable to expect the States participating in the State-NAEP program to provide or to assist in providing any information needed to prepare the sample of students. The level of resources provided by the Federal and State governments may determine how the National-NAEP procedures for sampling are applied in the individual States for the State-NAEP program. The actual sampling procedures must be consistent regardless of whether a single contractor is responsible for doing all of the work for drawing samples or whether States contribute substantially to the process. If necessary, States could do much of the clerical work involved in the sampling process and National-NAEP procedures could be followed to insure the integrity of the sample.

The sampling procedures followed in a State-NAEP program might be similar to those used by the National Assessment of Educational Progress in its project with eight Southern Regional Education Board States from 1985 to 1987. In this project a stratified random sample of schools from all regions and types of communities was selected in each State. Within each school a random sample of students was selected to participate. The results estimated achievement of all students at a given grade level in the

State. To permit simple estimation of standard errors, the sample was selected in the form of two sub-samples and was designed to achieve approximately 100 clusters of 20 students each. This resulted in roughly 2,000 students completing the assessment in each State. That sample allowed for State-level information by gender and race. From the background questions administered to students, it was possible to obtain information in a number of categories such as information on students enrolled in a particular curriculum in high school.

The sample of approximately 2,000 students per State per grade level would appear to be the minimum that could be used. The Southern Regional Education Board States' project typically tested students in 80 to 100 schools. The Alexander-James Report suggested testing approximately 4,500 students per State per grade level in about 90 schools. The size of the sample of the students to be tested has several factors which should be considered. First, sample size obviously should be sufficient for providing information on each population subgroup--for example, by race/ethnicity, gender, and type of community in which students live. Second, the number of students tested and the number of schools involved in the State-NAEP program must be sufficient to have face value and credibility with the public. It might be possible to obtain the State-level results by testing students in 50 schools or less in a State, but it might be impossible to convince the public that this was a satisfactory sampling of the State's schools. At the other end of this issue is a cost factor. Generally speaking, the number of schools involved in the sample has more of a bearing on the cost than does the number of students tested. Therefore, this argues for a realistic number of schools to be included in the sample. The number used by the Southern Regional Education Board States and suggested by the Alexander-James Report appears to be an appropriate number. Regardless of the sample size, the actual administration of tests should occur in groups of students of 30 or less. This size of testing group lends itself to better overall test administration than does a larger group.

States participating in the State-NAEP program should have the option of increasing the sample size in order to provide greater levels of detail. Options for increased sample size should be prepared for participating States. Increased sample sizes--be it in numbers of schools or numbers of students--involves increased costs. The increased costs for these optional sampling procedures should be paid by the States choosing the options.

While the sampling process is one in which States may provide assistance to the National-NAEP contractor or staff, the scoring and analysis of State-NAEP test booklets or answer sheets must be handled entirely by core staff under the direction of National-NAEP. Completed test booklets and answer sheets should be handled in the new State-NAEP program as they have been in the National-NAEP program and Southern Regional Education Board's project--that is, they should be sent to a central location directly from the schools immediately after completion of the testing program. In a fully operational State-NAEP program with several hundred-thousand completed test booklets or answer sheets, it may

be necessary to have regional processing centers or subcontracts for the processing of the answer booklets or sheets. Consideration should be given in the National-NAEP and State-NAEP program to conducting testing in an 8-week period as opposed to the current testing period which is an approximately 12-week period. The current plan for National-NAEP is to begin testing in January of even-numbered years. State- and National-NAEP should seek to complete testing shortly after March 1. The fact that National-NAEP has to "negotiate" with districts to get them to participate has apparently been one of the reasons for a 12-week testing period. This negotiation phase should not be a part of the State-NAEP program, for when a State agrees to participate in State-NAEP, it assumes the obligation of securing the participation of those schools chosen for the sample. The dynamics of the State-NAEP program could actually help the National-NAEP program because of the larger numbers of districts chosen for State-NAEP. In the past, very few school districts in a given State were chosen for the National-NAEP program. In a State-NAEP program approximately 200 to 300 schools per State could conceivably be chosen to participate. This could make schools less hesitant to agree to participate in National-NAEP because now they will be one of many schools in a NAEP program instead of one of a very few schools.

National-NAEP should be responsible for presenting the results of the State-NAEP program. These results should be presented in accordance with agreed-upon formats and the emphasis should be in presenting results. Interpretive comments should be provided by National-NAEP, but the initial emphasis should be on the presentation of results and the making available of these results for broad scale comment by educational and governmental leaders across the Nation. The initial emphasis in State-NAEP results should not be on report writing. A presentation of State-NAEP results should not be delayed while comprehensive reports to interpret the results are prepared.

National-NAEP should seek to make the release of State-NAEP results an "educational event." If the State-NAEP program is successful, the release of State-NAEP results could come to overshadow current educational events such as the release of the SAT and ACT score results and the release of the Secretary of Education's annual wall chart. When the State-NAEP program is fully operational, release of results should be a simultaneous event in State capitols and at National-NAEP headquarters. State educational and governmental leaders will provide interpretive comments on a State-by-State basis. The media in the respective States may be--or may be encouraged to be--interested in what the results mean for the individual States.

Those involved in the State-NAEP program should not lose sight of one of its most important values--that is the value of focusing public attention on educational improvement. The State-NAEP results could receive widespread attention because they will fill a void in educational information. Currently the SAT and ACT are the only tests that provide State-by-State information on a national basis. Generally speaking, the SAT and ACT results are the only student testing stories that make for front-page newspaper articles. Rarely do individual State testing programs receive the press coverage given to the SAT and ACT results. State-NAEP

programs could receive the same kind of widespread public attention and provide numerous opportunities for State educational and governmental leaders to focus public attention and raise important questions about education. For this reason, consideration should be given to releasing different subject-area results from the State-NAEP program at different times for the purpose of generating increased attention to results. If four subject areas are tested in the State-NAEP program in a given year, it is conceivable that the results from two of those programs could be released followed by somewhat later presentation of the results from the other two subject areas.

The present turnaround time of presenting results from the National-NAEP program is unacceptable for a fully operational State-NAEP program. Results from the State-NAEP program should be available in the same year that the testing is conducted, preferably in the early fall of the year if testing is conducted in January and February. At first look this schedule might appear impossible, given that the National-NAEP program now operates on an approximately 18-month turnaround schedule. That is, National-NAEP results are available approximately 18 months after testing begins. The State-NAEP program simply must be organized differently and have as an overriding objective to produce results in a timely fashion. If from the beginning a State-NAEP program is designed and organized to deliver in the fall results from a January-February testing cycle, it is reasonable to expect that this schedule can be met. Again, the emphasis must be on reporting results with some limited interpretation and allowing further interpretation to be made by various parties once the results are presented.

Two important questions that have to do with the management and administration of State-NAEP and also with its validity and credibility are: Who is tested? And who does the testing? The first question has major implications for test scores and their comparability. The answer to this question may have as much bearing on a State's overall scores as any single controllable factor. The second question has major implications for test scores, their credibility and comparability, and the program's cost.

Who is tested? The question "Who is tested?" and the reverse of that question--"Who is not tested?"--seems simple enough on the surface. The experience of 3 years of testing in the Southern Regional Education Board States' project with the National Assessment indicates that when these decisions are made, they are not as simple as it might appear. Once the sample of students is chosen to participate in State-NAEP, a decision must be made about whether the students can participate in a fairly routine testing program. Since many handicapped students are in regular classrooms today, this question becomes more difficult. The current National-NAEP guidelines call for excluding three categories of student. Those are non-English speaking students, educably mentally retarded students, and functionally disabled students--that is, students with temporary or permanent physical disability. The aim in the National-NAEP program has been to keep at the smallest possible level the number of excluded students. For example, when National-NAEP applied these guidelines to the 1984 and 1986 reading assessments at ages 9, 13, and 17, fewer than 4

students out of 100 were excluded in any year at any of the grade levels. State-NAEP should have a similar aim.

For State-NAEP, States should seek to follow an agreed-upon set of definitions for determining who is tested. Those definitions should be the ones used by National-NAEP. Before the 1990 State- and National-NAEP administrations, these definitions should be reviewed by States and the ETS/NAEP staff and refined where possible. One of the refinements needed is a more precise operational definition of students with limited English proficiency. The National Assessment Planning Project of the Council of Chief State School Officers has made a recommendation about how to define students with limited English proficiency. This revised definition should be considered by ETS/NAEP.

The application and interpretation of the definitions for excluding students from the testing program will still rely upon judgments by school personnel and/or test administrators. The National-NAEP practices for excluding students from assessments will likely differ from practices used by States on their own testing programs. The eight Southern Regional Education Board States demonstrated this clearly in their project with the National Assessment of Educational Progress. In a fully operational State-NAEP program, well over 10,000 schools could be involved. Thousands of persons would make decisions about which students would be tested and which would be excluded. These persons would be using guidelines that are not absolutely clear-cut. This is simply a fact of the State-NAEP program and not a weakness. It indicates that the issue of "who is tested" will have to be resolved and refined over a period of time.

That resolution and refinement will hinge in large measure on careful and comprehensive audits under National-NAEP auspices of the individual schools' lists of students to be tested, the number of students excluded, and the reasons for the exclusions. These audits should be conducted to determine if the definitions for "who is tested" are working and what corrections are necessary in the process or in the compilation of State scores. The percentage of students excluded in each State should be a part of the public release of the results from the State-NAEP program.

The State-NAEP procedures should make it difficult, or at least make it require a special effort, to exclude a student from testing. That required special effort would entail giving a written reason for each student excluded from testing. An audit of the school decisions about excluding students could signal any unusual or abnormal patterns and enable those who oversee the State-NAEP program to determine the reasons for these unusual patterns and whether they warranted corrective action or changes in policies.

The "who is tested" question would be central in the training of test administrators for State-NAEP. The administration of the State-NAEP test should not be a difficult or unusual procedure. Thus, a substantial amount of the training time could be allocated to the question of who is tested. The experience of the Southern Regional Education Board States' project was that over time, test administrators gained greater confidence in applying

the definitions for excluding students and the ETS/NAEP staff and the SREB States' test administrators expressed their views that the process improved with each test administration.

The second question that has major implications for test scores, their credibility and comparability, and the program's cost is "Who does the testing?" This may be the major question to be resolved about the management and administration of a State-NAEP program. There are two easy ways to resolve this issue, but neither of these appear to be practical. First the resources that the Federal and State governments pledge to the State-NAEP program could be sufficient to provide for the administration of the State-NAEP program in the same way as the National-NAEP program. That is, the National-NAEP contractor could replicate the procedures used in National-NAEP for the State-NAEP program. The problem is that the estimated cost of doing this is approximately $27 million. To date, there is no commitment by Federal and State governments for that level of support of the State-NAEP program.

The second easy answer to who does the testing is to allow individual States to be responsible for administering the State-NAEP test. State testing directors could argue that the State-NAEP test administration is not significantly different from the tests States give each year and that the State testing programs could efficiently administer the State-NAEP test. That may be true, but there does not appear at this time to be substantial sentiment for making the State-NAEP test administration a responsibility of the individual States. There appears to be a stronger sentiment that the State-NAEP administration requires a procedure that will have a high degree of public credibility and, therefore, a disinterested party must be involved in the test administration. A lack of dollars and a lack of credibility may therefore rule out the two easy answers to the question of who does the testing.

Test credibility should underlie decisions about State-NAEP test administration, but between a relatively expensive turn-key test administration with a national contractor and the States being entirely responsible for test administration, there are obviously several other options. Those options must weigh costs and relative levels of "disinterestedness" by the test administrators.

The logistics of the State-NAEP test administration also point to the fact that several options are possible. For example, at each grade level it is feasible to think in terms of 80 to 100 schools being involved in the State-NAEP program. It is also feasible to think in terms of a testing period that covers at least 8 weeks. Even an 8-week testing period would be substantially shorter than the current National-NAEP testing time period. Testing in 80 schools in an 8-week period would mean testing in 10 schools per week or two schools per school day. Devising a procedure to test in two schools per day in a State would seem to be a very manageable task even when testing at three grade levels is assumed--that is testing at six sites per school day in the entire State. A relatively small cadre of test administrators could seemingly handle this task. These testing administrators need not be highly paid Ph.D. psychometricians. Competent

persons with good managerial skills can direct the test administrations.

Another possible option is to use staff at community colleges which in most States are distributed across the State. These persons could form a cadre of test administrators. They would bring a level of disinterestedness to the process and they would be employees whose time could be contributed to the State-NAEP program. States participating in the State-NAEP program could be charged with the responsibility of recruiting a cadre of test administrators to work for and under the auspices of the National-NAEP contractor.

Whatever option might be used to bring a disinterested party into the classroom as a test administrator, there is an important change in the National-NAEP administration that should be considered for National-NAEP and State-NAEP. Specifically, the disinterested test administrator should be assisted by a school faculty or staff member in the testing process. The reasons for this are relatively simple but important. The administration of the State-NAEP test should occur in an orderly setting, and in a setting with 30 or fewer students. That setting is likely to be more orderly and reflect a more normal school situation if a school faculty or staff member is present in the room where the test is being conducted. That is not necessarily the case in the current National-NAEP test administration. It is possible that the National-NAEP tests are administered by a person whom students have never seen before and will never see again. For younger students this situation may produce anxiety and for older students it may result in a lack of motivation. Neither situation contributes toward an accurate reflection of students' achievement and performance. The new State-NAEP program should consider having a disinterested party in the classroom in charge of certain procedures such as timing, distribution of test booklets, and collection of test booklets and answer sheets. The school faculty or staff member should play a role in preparing students for the test administration, perhaps in reading the instructions of the administration and in general seeing that the process is an orderly one. This situation would more likely lead to student performance reflecting students' true abilities and achievement. The present National Assessment program may well be understating the achievement of our Nation's students particularly because of motivational factors for 13- and 17-year-olds.

Whatever options are chosen for administering State-NAEP, test administration manuals will be required for each grade level. Prototype State manuals have already been prepared by States including those prepared for the Southern Regional Education Board States' project. These manuals can best be prepared by States and the National-NAEP contractor working together. The experience of the SREB States was that the more States are involved in the preparation of the manuals, the better the finished product. National-NAEP might well want to subcontract for the preparation of the State-NAEP test manuals with the stipulation that the contractor seek broad-based input from States and State testing officials.

The administration of a State-NAEP program will require a financial commitment by States. Federal legislation may well require a State

contribution to the State-NAEP program. Even in the absence of such a Federal requirement, the cost realities of the State-NAEP program will mean that States must make a contribution. State government leaders who have been largely responsible for the efforts to establish a State-NAEP program will likely be willing to help underwrite the program. There will be practical limits, however, to what States will be willing to do. If test administration were to cost several hundred-thousand dollars in direct expenditures for each State, the participation of States would likely be reduced.

For additional information on the Southern Regional Education Board/National Assessment of Educational Progress project:

Southern Regional Education Board. (1984). Measuring educational progress in the south: Student achievement. Atlanta, GA: SREB.

Southern Regional Education Board. (1985). Measuring student achievement: Comparable test results for participating southern States, the south, and the Nation. Atalanta, GA: SREB.

Southern Regional Education Board. (1986). Measuring student achievement: Comparable test results for participating SREB States, the region, and the Nation. Atlanta, GA: SREB.

Southern Regional Education Board. (1987). Measuring student achievement: Comparable test results for participating SREB States, and the Nation. Atlanta, GA: SREB.

Recommendations for A Biennial National Educational Assessment, Reporting by State

R. Darrell Bock
University of Chicago and NORC

Except for the scale of operation, a national assessment of educational progress (NAEP) reporting at the State level is not fundamentally different from the present NAEP reporting at the national and regional level. But consideration of scale is important in the planning and execution of the design because the much greater costs will put a premium on returning a maximum amount of information for the expenditure. The present document outlines a series of recommendations for a State-reporting national assessment that will be cost-effective as an information system. Only issues of data collection and analysis are considered. Additional recommendations covering curriculum and standards for item and exercise construction are required to specify the assessment design fully.

The document is in three parts: Part I sets out general considerations leading to the main body of recommendations; Part II contains a numbered outline of the detailed recommendations; Part III clarifies and comments on each recommendation by number.

Part 1. General Considerations

1. Scale of operation and continuity

One implication of increased scale is justification for a larger initial outlay for planning procedures and development of the assessment and instruments than has been typical of NAEP in the past. The State-reporting assessment will, conservatively, yield 8 to 12 times as much information as the present regional-reporting NAEP and should therefore justify at least a four-to-sixfold increase in development funds. Especially critical is the enhancement of domain coverage and generalizability of assessment results through a greater number and variety of items and exercises in the cognitive instrument. The present NAEP instruments are too circumscribed to justify their application on the scale of a State-reporting assessment.

Another implication of increased scale is the greater effort that must be made to insure consistency of the assessment instruments, procedures and operations over years and decades and throughout the States and regions of the Nation. The potential monetary loss from technical errors that might invalidate an assessment is now so great that only the most deliberate and well-tested revisions and procedures can be introduced when changes in the instrument become necessary. The recent experience of procedural inconsistencies in the Reading assessment suggests that the present organization of NAEP, which lodges responsibility for design and procedures

with various grantees, cannot provide the required continuity. As the Alexander-James report recommends, responsibility for assessment policy, design and procedures should reside in a federally funded and staffed National Assessment Council or Center which will choose contractors to execute fully specified tasks of data collection and analysis. Only such an organization, modeled after other National Centers or Laboratories, can provide the procedural continuity that long-term measurement of educational progress requires.

2. Relative costs of sampling schools and students

A second consideration is that a large part of the continuing costs of a State-reporting assessment will be incurred in recruiting schools and sending assessment representatives to arrange and monitor the administration of the assessment instruments. Once the school has agreed to cooperate and local preparations for testing are complete, the marginal cost of adding more students to the sample within the school, or of obtaining additional data from each student, is relatively small. A fundamental principle of the assessment should therefore be that the quantity and quality of information returned from each school should justify the costs of reaching the school. This principle motivates the type of assessment design recommended in Section 2 and 3 of Part II. The design attempts to minimize the number of schools that must be recruited and visited in order to cover the cycle of subject-matter mandated in the pending legislation for the State-reporting assessment. The design also uses measurement techniques that will insure a level of data quality adequate for effective use of assessment results in educational policy, planning and research.

3. Updateability

"Updateability" is the provision for improving and adapting assessment procedures and instruments to changing conceptual and technological environments. Although its consideration is obviously in tension with that of continuity, both can be satisfied if certain principles of conservative management, already established in mature testing and assessment programs, are followed. They are the basis for the recommendations in Section 7, Part II, in which developmental research and field trials carried collaterally with the operational assessment are proposed. An orderly program of such study is required in order to insure compatibility of new and old procedures before changes are definitively introduced. To ensure continuity of the assessment instruments, especially the cognitive tests, the principle of strict specification of the structure of items and exercises within the test booklets is recommended in Section 3, Part II. This degree of specification is necessary to insure that item replacements do not alter the domain coverage within subject matters. Finally, scaling of the assessment results by means of suitable item response theoretic (IRT) models, already introduced by ETS, is essential in order to preserve the comparability of the reported attainment measures as items are periodically replaced to permit their public release and, where possible, to improve the psychometric properties of the instruments.

4. Reporting by student proficiency levels

The fourth consideration is that the assessment instrument be capable of estimating, with good accuracy, the locations of individual students on the main proficiency scales in terms of which the assessment results are reported. This requirement must be met in order to report the proportion of students at each grade level who exceed defined levels of performance on these scales. This type of reporting was not possible with the ECS assessment design, and it is possible in the ETS design only indirectly and with score attributions. In contrast, the type of instrument recommended in Part II, Section 3, is capable of sufficiently accurate estimation of each student's location on the scale to support direct reporting of the proportions of students in each performance category.

5. Performance anchoring of proficiency scales

A fifth consideration is that of interpretation of the assessment scale. In order to avoid purely normative interpretation, the critical points on the main proficiency scales should be "anchored" by empirical studies that establish 80 percent probability of successful student performance on realistic tasks related to the subject matter. An example of an anchoring study might be the following. A subsample of students who had participated in the group-testing assessment might be asked, in an individual testing session, to find a certain item in a mail-order catalogue and phone in and order for it to a (simulated) operator. Aggregated over a subsample of schools in several States, the results from this individual test could be used to estimate the probability of successful performance as a function of a student's location on the reading proficiency scale. This function would then serve to define one of a number of similar anchor points on the proficiency scale. Provision for performance anchoring studies is recommended in Section 7, Part II.

In order to provide for objective and practical interpretation of the proficiency scales, it is important that they are defined at each of the three grade levels in the assessment (4, 8 and 12) and not, as in the current practice, defined vertically over the corresponding wide age range. Not only is such vertical equating virtually meaningless in a subject such as mathematics, the content of which changes greatly over this range, but the proficiency levels on such scales are too widely spaced to be useful. Most fourth graders are necessarily at the lower levels of the scales and most twelfth graders at the higher levels purely for developmental reasons unrelated to the effectiveness of instruction.

6. Provision for multilevel analysis

A consideration that has influenced the recommended design, and departs from past NAEP procedures, is the provision for multilevel analysis of the assessment data. Although not reported directly, the estimated locations of individual students on the main proficiency scales should be available for statistical analysis by the National Assessment Center, the National Center for Education Statistics, the departments of education of the participating States, and qualified educational research workers.

Moreover, scores at the school level should be available both for the main proficiencies and for detailed curricular objectives that can be assessed only in group data. NAEP should employ the school as a unit of analysis and reporting, and the assessment design should provide for efficient estimation of school-level variation. To encourage the participation of schools, summary reports of a school's performance comparison with national norms should be made available in a timely fashion to those schools that request them. At the State and national level, distributions of school means and other school-level statistics should be part of the assessment reports in addition to the social group reporting typical of past national assessments.

7. Interpretation

A final general consideration is that, in order to ensure prompt release of assessment results, interpretation should be separated from the reporting of current figures and their statistical significance. A statistical report should be released early in September of the assessment year. It should present summary and detailed information in tabular and graphic form, and should certify the statistical significance of trends and differences. More studied interpretations and policy recommendations can be deferred to a later time and relegated to other reports.

Part 2. Detailed Recommendations

1. Sample

1.1 Populations in participating States.

All public school children assigned to respective grades 4, 8 and 12 at the time of testing (February) will define the populations for sampling purposes.

1.2 Sample for State-level reporting.

Up to sixty in-scope students sampled randomly from each of the three grade levels in each of 80 to 100 schools of a stratified sample of schools in each participating State. Sample stratification design should facilitate within-State comparison of major school types (urban, rural, other) and major ethnic groups. No school shall appear in the sample more frequently than every 4 years.

1.3 Scope definition.

A uniform definition of in-scope students and a procedure for local identification of out-of-scope students should be based on guidelines from the Council of Chief State School Officers.

1.4 State "buy-in" sample augmentations

States may purchase, at cost, sampling of any number of additional schools to provide reports for designated subpopulations within the State.

2. Subject-Matter Allocation

2.1 The assessment cycle.

Within each school, two groups of up to 30 students will be tested independently in the following fixed cycles of subject matters:

		Year of 4-Year Cycle A		B
Group 1	Test a:	Mathematics	Test a:	Mathematics
	Test b:	Physical, Biological, & Earth Sciences	Test b:	Social Science & Civics
Group 2	Test c:	Reading in Science, Literature, & Social Studies	Test c:	Reading in Science, Literature, & Social Studies
	Test d:	History, Geography, & Literature	Test d:	Writing & Language Arts

2.2 Domain definitions

Domain definitions, including specification of content and process categories, will be available to the public at least 1 year in advance of testing in each of the subject matters. Domain definitions will be revised in not less than 12 years and not more than 8 years from the most recent revision.

3. Assessment Instruments

3.1 Item formats (for all subject matters except writing).

3.1.1 Brief-answer free response

Each student booklet of the assessment instrument will contain 5 brief-answer free response items.

3.1.2 Multiple-choice

Multiple-choice items will contain not less than 4 alternative answers. Examinees will be instructed not to make blind guesses if they do not know the answer, but to go on to the next item.

3.2 Sources of items

In addition to items written expressly for the State-reporting assessment, items should be sought from the existing NAEP item pools and from those of the State assessment programs and international studies of school achievement. Items from State assessments should not appear on State instruments that might be administered to students in the sample for the State-reporting assessment of the same year.

3.3 Configuration of the cognitive instrument for all subject matters except writing

Two-stage, multiple-matrix sampled instrument structured in each subject-matter area follows:

Pretest: 16 items plus student background questionnaire (see 3.4). Neither the pretest nor the questionnaire is matrix sampled.

Variant items: Two experimental items will follow the pretest items on each questionnaire. These items will be used for instrument updating (see Section 7.1).

Second-stage test: 48 to 64 items per booklet

		Forms							
		1	2	3	4	5	6	7	8
Booklet within forms	1. "Easy"	+	+	+	+	+	+	+	+
	2. "Intermediate"	+	+	+	+	+	+	+	+
	3. "Hard"	+	+	+	+	+	+	+	+

Booklets within forms are linked by 32 common items per form. Number of distinct items per form is 112. Number distinct items required, including the pretest, is 912. There are no common items between forms.

3.4 Writing and language arts

The writing and language arts test contains two sections:

Section 1

Twenty objective items covering mechanics of writing (spelling, punctuation, capitalization, grammar). A total of 480 distinct items are assigned to 24 test booklets of the matrix sample.

Section 2

Prompts for the writing test are classified by type of writing (e.g., autobiographical incident, evaluation, story, etc.) Six prompts for each of four types are randomly assigned to the test booklets, one prompt per booklet. Each student writes a brief essay in response to the prompt in his or her booklet. These booklets also contain the 20 objective items on mechanics of writing.

3.5 Content by process structure of the test booklets for subject matters other than writing

Items within each test booklet are classified by subject-matter content and cognitive process. In so far as possible, each booklet contains one and only one item representing each intersection of the content and process categories. Each test booklet replicates the item structure except for position within the booklet: order of the content by process subclasses identical with forms but rotates from one form to another.

3.6 Background questionnaire

The assessment instrument includes the following questionnaires. The lengths of the questionnaires should be adjusted so that the times required for completion are as indicated.

1. Student: 20 minutes plus 20 minute pretest (see 3.1).

2. Teacher(s): one hour.

3. Principal: one hour.

4. District Superintendent: one hour.

Some items of the student questionnaires will be specific to the subject matter of the assessment test that the student will be administered.

Only one form of each of these questionnaires will be used as appropriate for the subject matter of a particular assessment. Item content should be coordinated with other educational surveys such as High School and Beyond (HSB), the National Educational Longitudinal Study (NELS) and the studies of the International Educational Achievement Association (IEA).

4. Administration of the Assessment Instrument

4.1 Testing conditions

Students will be tested in groups of not more than 30 in a quiet classroom reserved for that purpose.

4.2 Assessment representatives

An assessment representative will be present during all testing sessions and will be responsible for 1) distribution and collection of the booklets, pencils and scratch paper, 2) timing of the test administrations, 3) dispatch of the completed test booklets to the national or regional center, and 4) security of all restricted materials.

4.3 Test administrators

Explanation of the testing and reading of the test instructions will be presented by a teacher or counselor assigned for this purpose by the school principal. The test administrator will be instructed by the assessment representative.

4.4 Assignment of the assessment representatives

Each State will provide the services of 12 to 14 assessment representatives for a 3 day training period and 25 days at the school sites. In so far as possible, the representatives should come from communities in the State that will minimize the need for overnight travel to schools. Representatives should not be regular employees of the schools they will monitor. The national assessment center will be responsible for training the assessment representative.

4.5 Testing schedule

Student time for testing is 2 hours divided between a morning and afternoon session. An additional hour for orientation, questionnaire completion and pretest is required on a day preceding the main testing. The assessment representative must be present on both days. The hourly schedule, which must be the same for schools in all States, is as follows:

Day 1		Day 2	
10:00- 11:00 am	Group 1 pretest & question- naire	10:00- 11:00 am	Group 1 Test a
11:00 am- 12:00 noon	Group 2 pretest & question- naire	11:00 am- 12:00 noon	Group 2 Test c
		1:00- 2:00 pm	Group 1 Test b
		2:00- 3:00 pm	Group 2 Test d

4.6 Scoring the pretest

During the afternoon of Day 1, the assessment representative will score the pretest and assign the second stage tests to the students accordingly. At the same time, the representative will check that the name field, sex and birth date entries are complete and have them corrected if necessary.

4.7 Absent students

Ten alternate students will be selected with the sample. These alternates may replace any student who is absent from testing for reasons clearly unrelated to attainment (e.g., verified illness, accident not the fault of the student, death in the family, etc.) Students absent on the first day only will be administered the questionnaire and pretest during the first class hour of the second-stage testing day. The assessment representative should arrange to call back to the school to test any other absent students on a later day.

4.8 Dates of testing

All testing should be completed in the first 4 full weeks after the 31st of January. Up to five additional days can be used for call backs. Regular test days should be Mondays through Thursdays; Fridays should be reserved for call backs.

5. Scoring the Tests

5.1 Brief-answer free response tests

Readers should rate free-response items for quality or degree of correctness on a scale with up to 7 ordered categories, as appropriate to the particular items. Rating protocols for each subject-matter area will be supplied by the national assessment center, which will also train the readers. Readers can be engaged by the States, but their activities must be monitored by the national assessment center.

5.2 Multiple-choice tests

5.2.1 Creation of the student-level response file

After scanning and cleaning of responses to multiple-choice items, the pretest, questionnaire, and second-stage item response are merged into the student-level master file. Each record will include the identification of the second-stage form and the assessment identification code of the student and school. A roster containing these code numbers will be supplied to the district superintendent, or school director, in printed and machine-readable form.

5.2.2 Student-level scaling of main content and process dimensions

Student-level attainment values on defined content and process proficiency scales within each subject-matter will be computed from each booklet using a suitable and well-fitting item response model. The model must incorporate information from first- and second-stage item responses and account for effects of guessing. Standard errors for estimated student attainment values will not exceed one-half standard deviation of the national distribution for the corresponding scale.

5.2.3 School-level scoring of detailed content by process categories

A school-level score for each content-by-process intersection in each test will be computed by aggregating responses across forms and using a suitable and well-fitting group-level item response model incorporating pretest information and accounting for guessing.

5.2.4 Standard NAEP scales

The score for each scale will be reported as the true score on a formally defined "ideal" test consisting of 500 items conforming to a one-parameter logistic item response

model with location parameters positioned at equal intervals between -5.0 and 5.0 standard deviations of the latent score distribution, and a slope parameter equal to the geometric mean of the slopes of items of the actual scale. For reporting purposes, these scores will be referred to as "scale" scores.

5.2.5 The origin and unit of the scores in each scale

The origin and unit of the scores on each scale will be set to 250 and 50, respectively, in the national sample distribution in the first year of the assessment. The origin and unit will be maintained throughout the tenure of the scale in the assessment instrument (see Section 7.1 on instrument maintenance.)

5.3 Writing test

5.3.1 Reading and grading

Student essays will be read by a member of a writing team trained in the grading of one of the types of writing. The reader will rate the essay on several six-point scales with defined ordered levels of attainment. Four to eight scales measuring distinct aspects of effective writing will be defined. (See Section 7 on maintenance of the essay grading procedure.) Readers may be engaged by the States, but their activities must be monitored by the national assessment center.

5.3.2 School-level scaling of the ratings

Essay ratings will be scaled using a school-level item response model for ordered categories common to all prompts within writing type. The model will include a location and slope parameter for each item, and parameters for the internal boundaries of the common categories. School-level scale scores having the same properties as those of the multiple choice tests will be computed for each scale and writing type. Mean school scores over writing types will be computed with standard errors reflecting the sampling of prompts within writing types, and the sampling of students within schools.

5.3.3 Student-level information

For some purposes of secondary analysis, student-level scores will be assigned to ratings as midpoints of the score category, adjusted for the location parameter of the prompt.

5.4 Comparable scale units at student-level and school-level.

For the free-response and multiple-choice tests, the unit of scale of the school-level scores is set in first assessment year so that the residual variance from the regression of each scale on school background characteristics is equal to the residual variance from similar regressions of the school means of student-level scores, in the same content area and within subject matters.

5.5 Aggregation of student-level and school-level information.

Student-level scores are aggregated without weighting to obtain school means, standard deviations and other school-level statistics. The same scores are aggregated with State sampling weights to obtain State means standard deviations, percentiles, and other State-level statistics. State statistics are aggregated to national means, weighted by numbers of students in the State at the respective grade levels. School-level scores are aggregated to the State level, weighted by the sums of the student-level weights of the school at the respective grade levels.

6. Reporting

6.1 Performance anchoring of proficiency scales

For each subject-matter except writing, relevant performance tests will be administered to a 20 percent random sample of students taking the assessment tests in 10 schools in each participating State. Data from this collateral testing will be used to determine "anchor" points on the corresponding subject-matter assessment scales where 80 percent of students nationally exceed defined performance levels (see Section 7.2).

6.2 Standard errors

All scale scores for individual students will be reported with standard errors reflecting the sampling of items for the test booklets. All summary scores for school, State and Nation will be reported with standard errors reflecting the sampling of schools, students, and items.

6.3 Report to the National Center for Education Statistics

On the day after Labor Day of the assessment year, reports of means, confidence intervals, and distributions of student- and school-level scores on the attainment scales will be delivered to NCES in tabular and graphic form. Information concerning background data for students, teachers and schools and their relationship to the attainment data will be included. Separate results for each participating State and a national summary will be shown. Comparisons for previous years will be presented in tabular and graphic form.

6.4 Reports to State departments of education and to schools

On or about September 15 of the assessment year, State departments of education in participating States, and district superintendents of participating schools will be sent a report of content by process scale-score distributions for the samples of students tested in the schools. These scores will be referred to percentiles of the State and national distributions. Individual students will not be identified in these reports. School-level scores for detailed content by process variables will also be reported.

6.5 Data for linking the national assessment scales to those of State assessment programs

Files containing student- and school-level scale scores, and questionnaires data, identified by codes known to the respective district superintendents or school directors, will be delivered to the State departments of education in January of the year following each assessment. These files are intended for use in linking national assessment scales to those of the assessment programs in participating States (see Section 7.4).

6.6 Secondary user files

Files containing student item responses, student- and school-level scale scores, and questionnaire responses, with non-informative student or school identification, will be released to secondary users in January of the year following each assessment. States will be identified in these files.

7. Technical Support

7.1 Assessment instrument maintenance

7.1.1 Free response and multiple-choice tests

A 5 percent probability sample of the national assessment data will be examined biennially for evidence of drift in item locations on the assessment scales. Parameter values in the psychometric scaling models will be adjusted if necessary or items will be marked for replacement. Ten percent of free-response and multiple-choice items will be replaced with new items in each assessment. The variant items appearing with the questionnaire and pretest of a previous assessment may be among those used as replacements (see Section 3.3.).

7.1.2 The writing test

Twenty-five essays selected randomly for the two most recent writing assessments will be randomly seeded among

the essays read by each reader in the current assessment. The origin and unit of each reader's scale scores will be adjusted so that the mean and standard deviation of the reader's scores on the seeded essays will equal their previous values. Each reader must read approximately equal numbers of essays written to each of the prompts of scales scored by that reader.

Readers who are excessively deviant, and prompts that are unproductive or inconsistently read, will be marked for replacement. Twenty-five percent of essay prompts will be replaced with new prompts in each assessment. Scale scores of old prompts in the current assessment will be analyzed in conjunction with scores based on re-reading of previous essays to estimate and adjust for effects of exposure of the prompts.

7.1.3 Questionnaire

Questionnaire items will be reviewed at 4-year intervals for productivity and relevance. Unsuitable items will be replaced.

7.2 Performance test development

A performance test development team will function collaterally with the assessment operation to study, prepare, analyze, and train administrators of the performance tests used in anchoring the multiple-choice assessment scales.

7.3 Scale validity and item writing technique

A team to investigate the construct and external validity of the assessment scales, and to evaluate techniques used in item writing, will function collaterally with the assessment operation.

7.4 Linking of national assessment scales to those of assessment programs in participating States

The national assessment will supply, to State departments of education, data files of State results in a form suitable for predicting national scale values from data of the State assessment program. The national assessment will provide States with computerized statistical procedures for equating State and national results.

7.5 Methodological development

A team functioning collaterally with the assessment operation will investigate and develop new methods for the conduct of large-scale assessment programs. Initially, priority will be given to 1) character reading or computerized response modes to permit the use

of free-response items in the assessment instrument, 2) investigation of panel effects when the same schools are tested in two or more assessments, and 3) sampling designs for best joint estimation of student effects and school effects.

Part 3. Comments and Clarification

1. Sample

1.1 Populations in participating States

A strict grade assignment definition of the populations is recommended in order to bring the assessment results into a closer relationship with the instructional programs of the States. The age of the students will be known from the background questionnaires and can be used in accounting for student attainment in States with differing rules concerning age of school entry.

The exclusion of private schools from the State-reporting assessment may present serious difficulties for between-State comparisons. Statistical adjustments to account for the varying private school populations may be required for interpretation of differences between States. Data on private schools from the National-NAEP or other surveys will be needed in this connection.

1.2 Sample for State-level reporting

The primary sampling unit will be schools drawn from lists of public schools provided by the State departments of education. The sample should be sufficiently large to provide reasonably accurate estimation of within-State comparison of major school types, classified as "urban," "rural," and "other," and of major ethnic groups when sufficiently represented within the State.

If the same school is selected for the National-NAEP and the State reporting NAEP, the National-NAEP should have priority. No school should appear in the sample of either of the assessments more frequently than every 4 years.

Sampling of students within the schools will be carried out by the National Assessment Center from rosters provided by the selected schools. Number of schools in the sample and the number of students per school will be determined so as to minimize the cost of obtaining a specified standard error for the estimated State mean.

The definition of which students are untestable and thus out of scope for the assessment must be defined at the national level and applied uniformly in all States. Guidelines set by the Council of Chief State School Officers should be followed in identifying students unable to take the test for reasons of physical or psychological limitations, or inability to read English.

1.3 State "buy-in" sample augmentations

States should be permitted to pay for the sampling of additional schools up to and including a census of all schools in the State, if they so desire. In the latter case, a unit of the assessment dedicated to the State would probably have to be organized, but this level of testing should perhaps be postponed until the basic State-reporting assessment is in stable operation.

2. Subject-Matter Allocation

2.1 The assessment cycle

The pending legislation proposes a biennial assessment in which Mathematics and Reading appear each year, Science and History, Geography and Literature every 4 years, and direct Writing assessment every 6 years. These are minimum requirements. The cycle recommended in 2.1 exceeds these requirements in that all of the subject matters are brought within a 4-year cycle. As mentioned in the introduction, in order to maximize cost-benefit from recruiting and visiting schools, the information return is increased by testing two independent groups of students in two of the assessment subject matters, one of which is either mathematics or reading, which must appear in each assessment. In addition, each of the groups is tested in a second subject matter that differs in the "A" and "B" years of the 4-year assessment cycle. This allocation scheme covers all of the required subject matters in the fewest feasible school visits.

2.2 Domain definitions

Within each of the six subject matters of the assessment, it is recommended that a detailed domain specification, based on a content by process classification of curricular objectives, be released publicly well before the testing in each area. Comment and suggestions on the definitions should be received and taken into account before the final version of the assessment instrument is completed and adopted. It is not expected that these definitions will change at every assessment. Indeed, it is recommended that, in order to protect the assessment from the short-lived fads and enthusiasms characteristic of the educational field, definitions not be revised more often than every 8 years. Nevertheless, to provide for genuine progress in the domain definitions, revisions should be required at least every 12 years.

3. Assessment Instruments

3.1 Item formats

3.1.1 Brief-answer free-response

To prepare the way for reduced reliance on multiple-choice items as technology permits, each booklet of the cognitive assessment should contain five brief-answer free-response items. These types of items are most useful in measuring higher-order reasoning processes, where the response may

reveal not just correct or incorrect knowledge, but the degree of understanding or originality.

3.1.2 Multiple-choice

To control effects of guessing, the multiple-choice items should not employ less than four alternative answers, one of which is unequivocally correct. It is not recommended that the alternative "none of the above" be used as an additional option. Examinees should be instructed not to guess but to go on to the next item if they do not know the answer. In this way, the student has the opportunity to go back to an omitted item if there is time during the testing session.

3.2 Sources of items

In order for the assessment in each subject-matter area to represent the domain adequately, considerable numbers of items may be required. It is not necessary, or even desirable, that all of these items should be written expressly for the State-reporting assessment. Many suitable items already exist in the item pools of the State assessment programs and other testing efforts. Large numbers of items can be donated by the States, classified and entered into an information system that would make them available for instrument development. The same items should not, of course, appear simultaneously on State instrument and national instruments that might be taken by the same students during an assessment year.

3.3 Configuration of the assessment instrument

It is proposed that the cognitive assessment instrument be structured as a two-stage test, with a 16-item first-stage test to be administered along with the student background questionnaire on the school day preceding the second-stage test. The first-stage test will be scored with the aid of a scoring template by the assessment representative for the school, and an "easy," "intermediate," or "hard" second-stage test assigned to each of the selected students accordingly. Procedures for this purpose have been extensively investigated in Illinois and California, and have been found to improve substantially the quality of data obtained from assessment tests. (Reports of these studies are available from the Center for Research in Evaluation, Standards and Student Testing (CRESST), School of Education, UCLA.)

To provide a high level of generalizability of the domain mean scores, it is recommended that at least eight forms of the assessment test in each subject matter area, each consisting of an easy, intermediate and hard test booklet (a total of 24 booklets), be constructed in each subject matter area. Booklets within forms are linked by common items, but are not so linked between forms. Because the forms are assigned in random rotation to each second stage group, the common population method provides for equating of forms. Booklets are scaled by means of an item response model that includes the common and unique items.

If 48 items can appear in each second-stage test booklet, for example, each of the 8 forms might then have 112, unique distinct items and 32 common or linking items. The number of distinct items required for the instrument, including the pretest, would therefore be 912.

It is also recommended that two experimental items, called "variant" items, be included as items 17 and 18 of the pretest. These items will be different on each of a number of pretest booklets, which in all other respects will be identical. The purpose of including these items is to provide estimates of their psychometric characteristics as a basis for replacing items retired in future assessments.

3.4 Writing and language arts

Although productive writing proficiency can only be tested by some form of written essay, knowledge of the mechanics of writing, including spelling, punctuation, capitalization and grammar are better and more economically tested by objective items. The writing and language arts test should therefore consist of two sections: 1) a brief objective test of twenty items matrix-sampled from a total of 480 distinct items and assigned to each of 24 test booklets, and 2) a prompt for the writing exercise. The first section, which might take the form of a proofreading exercise, should be separately timed, require no more than 10 minutes, and should precede the writing exercise.

The students should write in response to prompts sampled from four such types of writing as autobiographical incident, report, evaluation, and story. Six prompts in each of these four types of writing, all previously tested for productivity, can then be randomly assigned to the test booklets. Each student should have 40 minutes to write a brief essay in response to one of the prompts. These recommendations are patterned after the California Direct Writing Assessment.

3.5 Content-by-process structure of the free response and multiple choice test booklets

In each of the subject-matter areas except writing and language arts, items within each test booklet should be random replications of a fixed subject-matter content by cognitive process classification. The main content and process categories define dimensions for a joint estimation of content and process proficiencies based on responses within each student booklet. In addition, each content by process intersect or subclass of the two-way classification, can be scored at the school- by aggregating responses across test booklets to items in each the intersections. This method of structuring assessment instruments, studied by Bock and Mislevy , provides both for assessing proficiency in main content and process dimensions at the individual student level, and for assessing attainment in detailed topics and skills in the curricular objectives at the school level. Both types of information are needed for the widest use of the assessment data. (See Bock and Mislevy, 1988)

3.6 Background questionnaire

The assessment instrument includes several questionnaires for obtaining background and other noncognitive information about students, teachers, and the school. The student questionnaire precedes the pretest in the pretest booklet. Each booklet contains exactly the same questionnaire items, but the questions may refer to the specific subject matter of the assessment. The pretest items are also specialized through the subject matter areas that the student will take in the second stage testing.

Teacher, principal and district superintendent questionnaires contain items intended to elicit information about the school resources and programs. The content of these questions should be coordinated with other national surveys of school characteristics.

4. Administration of the Assessment Instrument

4.1 Testing conditions

Testing conditions should conform to those typically required for achievement testing--freedom from distraction, good lighting, comfortable seating and adequate space between students.

4.2 & 4.3 Assessment representatives and test administrators

It is recommended that the administration of the assessment tests be conducted by two persons: 1) an assessment representative hired by the State testing director, but trained by the National Assessment Center, and 2) a test administrator assigned by the principal of the school. The assessment representative will instruct the test administrator and handle procedural steps in the distribution of booklets, timing of the administrations, and will ensure that acceptable conditions of testing and security for all test materials are maintained. The purpose of using a local test administrator is to facilitate communication with the students in the orientation and reading of instructions and also to avoid the well-known "substitute-teacher" effect when an unfamiliar person takes charge of a group of students, especially younger students.

4.4 Assignment of the assessment representatives

Testing director of participating States should arrange for 12 to 14 capable workers to spend approximately 6 weeks, beginning the last week of January and ending the first week of March, coordinating the monitoring of testing in the schools. One source of these representatives might be residents of the States who are part-time field workers of National Survey organizations. Another might be professional substitute teachers. They should be chosen from communities strategically situated in the State to minimize travel and overnight stays. Travel and accommodations within the State will be necessary, however, for a 3- or 4-day training period to precede the testing. These sessions will be conducted by personnel of the National Assessment Center.

4.5 Testing schedule

Testing at each school extends over 2 days: on the morning of the first day, two 1-hour sessions are devoted to the orientation of the selected students, administration of the student background questionnaire, and administration of the pretest. Half of the students selected in the school are tested in each of these sessions. The two groups take different second-stage tests on the next day, and so receive different pretests on the first day. The questionnaire, however, is common for all students apart from the variant items and questions relating to the subject matter.

On the second day, the two groups assemble once in the morning and once in the afternoon for 1-hour periods devoted to the second-stage test. Fifty minutes must be available in these hours for students actually to work on the tests. If there are any delays, these sessions must go overtime.

4.6 Scoring the pretest

During the afternoon of the first day, the assessment representative will score the pretest using a stencil provided for that purpose. In no case will more than 60 tests need to be scored, so that the task can be completed in about 2 hours Representatives should also check that essential information on the questionnaire has been coded by the student. If not, the students can be contacted the first hour of the second day for corrections.

4.7 Absent students

It is permissible to make use of alternate students, chosen at the time the sample of students is drawn, as substitutes for absent students, provided the reason for the absence is not related to school attainment. A list of admissible reasons for absence with substitution should be provided to the Assessment Representatives. For those student absent without an admissible excuse, every effort should be made to arrange retesting on a later day. For this purpose, the pretest can be given during the first hour of the day so that testing can be completed in 1 day.

4.8 Dates of testing

It is important that dates and times of testing coincide as closely as possible between States. Assuming a maximum of 100 schools tested per State and the availability of 14 Assessment Representatives, the testing could be completed in a four-week period with each representative testing two schools per week, preferably on Monday/Tuesdays and Wednesday/ Thursdays. Friday can then be reserved for callbacks to test absent students. On this schedule, all schools could be tested in a 4-week period during February. These dates of testing would interfere least with the State testing programs and other scheduled activities of students during the Spring.

5. Scoring the Tests

5.1 Brief answer free-response tests

The brief-answer items will have to be scored by readers trained for this purpose. State Testing Directors should retain persons for this purpose. Assuming five such items per test booklet and a State sample of at most 6.000 students, 12 readers could easily complete this work in 10 working days. The readers would have to be trained and the quality of their work checked by the National Assessment Center. It is highly desirable that brief-answer questions elicit responses that can be graded by quality or degree of correctness. Such items are rated on a seven-category ordinal scale. Each item will then convey approximately the same information as six multiple-choice items.

5.2 Multiple-choice tests

The scoring of the multiple-choice tests is almost entirely automated and can proceed quickly once the computer procedures are fully developed. The necessary computer programs are available either from the existing NAEP operation, from tate assessment programs, or by license from commercial test scoring companies. The statistical and psychometric procedures for scoring the two-stage tests are now well documented and will not be commented on in detail here. The main innovation of this section is 1) provision for conjoint scoring of the content and process dimensions of the cognitive instrument by means of a suitable item response model (Section 5.2.2), and 2) scoring for detailed curricular objectives by means of a group level item response model that aggregates information across test booklets. (Section 5.2.3.)

It is recommended that the scale scores computed using such models be defined in the manner of the present NAEP proficiency scales. That is, when the test is initially introduced, the mean will be set to 250 and the standard deviation to 50 in the national student population at the respective grade level. (Sections 5.2.4. and 5.2.5.)

5.3 The writing test

As discussed in Section 3.3, the writing test has an objective section on mechanics of writing that will be scored by machine methods. The second section, however, is a student essay that must be graded by human readers. Teams of readers must be organized and trained for this purpose. In the California Direct Reading Assessment, for example, teams of approximately 50 readers are trained in the reading of each of the distinct types of writing. Inasmuch as able readers can easily read and grade 60 essays per day, and only 3,000 students per State participate in the reading assessment, five readers in each State could complete the grading of the writing exercises in 10 working days. These readers could be retained by the State testing director, but trained and checked for quality by the National Assessment Center. Maintaining the consistency of grading from one assessment to the other by randomly including for re-reading essays from previous assessments is discussed in Section 7.

Two methods of scoring the writing exercises are recommended in Sections 5.3.2 and 5.3.3. For the California Direct Writing Assessment, an item-response model has been developed for scoring the graded responses at the school level by aggregating over the matrix sample the prompts in the test booklets. This type of scoring allows for replacement of a proportion of the prompts in each assessment in order to prevent their over-exposure nationally and to permit release of some prompts and writing examples for public discussion.

In addition, because of the graded scoring of the essays, values can be assigned to the score categories that make it possible to employ the response to the single prompt for the analysis of data at the individual student level.

5.4 Comparable scale units at the student level and at the school level

An essential feature of the scoring procedures proposed here is that proficiencies in the main content and process categories are estimated for individual students and subsequently aggregated to the school, State, and national level. At the same time, attainment in detailed subclasses of the content by process classification is scored at the school level, directly aggregating responses to items in that subclass across test booklets. These types of scores, which are essential for curriculum decisions and guidance of instruction, are then also aggregated to the State and national level. To simplify reporting, these two types of information should be expressed in scales with the same origin and unit of measurement. The most straightforward way of providing the equating of scale is first to set the student-level scales to a mean of 250 and standard deviation of 50 in the national population, then to adjust the mean of the school-level scales to equal that of the individual-level scales, and finally to adjust the unit of measurement so that the residual variance of the model predicting school performance from background characteristics is equal for the two types of scale. These equatings are performed for each content area within the various subject matters.

5.5 Aggregation of student-level and school-level information

Because students within schools are chosen by simple random sampling, student-level information may be aggregated without weighting to obtain means and standard deviations for the schools. The schools, however, will typically be selected into stratified samples in order to insure better estimation for certain subpopulations within the State, and thus require weighting of student-level data to estimate the State average. Case weights for this purpose should be assigned to each student-level data record. School-level scores can then be aggregated to the State-level weighting by the sum of the student-level weights in that school at the grade level. All State-level statistics are then correctly weighted and can be aggregated to the national level, weighting by the numbers of the students in the State at the grade level in question.

6. Reporting

6.1 Performance anchoring of proficiency scales

As mentioned in the introduction, the interpretability and practical relevance of the assessment scales will be greatly enhanced by a collateral program of performance testing carried out by a special team from the National Center. Members of the team will select schools from the assessment and carry out individual performance tests on a subsample of students who have previously taken the regular assessment tests. In performance testing, a so-called "work sample" analysis is used to define tasks that are in closer correspondence to real-world situations than are typical multiple-choice items.

Because performance tests require actual production from the student and not merely choices among preconceived alternatives, the administration and scoring of the test can be time consuming and expensive. The cost can be justified, however, because the program will demonstrate to the school the importance placed on evaluation of performance and products and not merely on the recognition skills tested by multiple-choice items. At the same time, performance testing provides the objective criteria by which the proficiency scales from the assessment (except for the reading test which is already a performance measure) can be given a concrete meaning. For each performance task, the probability of success can be expressed as a function of the student's location on the proficiency scale from the main assessment. The location on the scale where this function reaches 80 percent then defines an anchor point on the assessment scale. Any student located above that point has an 80 percent probability or better of passing the performance test.

6.2 Standard errors

Although it is not always made explicit, the procedures for a sampling assessment involve a three-stage sampling scheme. First, the schools are sampled, possibly in a stratified manner according to school characteristics, then students are sampled within grade levels, and finally an assessment booklet that samples items from the domain is assigned randomly to the student. Because the inferences to be drawn from the assessment data refer to all three of these populations--the schools, the students, and the item domains--it is important that a fully specified error model be employed when estimating standard errors for State and national means and other statistics.

6.3 Report to the National Center for Education Statistics

A much greater effort than in past assessments must be made for timely delivery of State and national results. A statistical report, including point estimates of all relevant statistics, confidence intervals, and estimated distributions for scores both at the student and school level, should be delivered to NCES by the day following Labor Day of the assessment year. If all scoring and data processing procedures are in place, the 6 months between the completion of testing the first week of

March and the first week of September would be sufficient to deliver a statistical report. State assessment programs based on a census of every student at several grade levels are able to deliver results in this time frame or better. It should also be possible for the national assessment, especially if more elaborate interpretative reports are postponed to a later date.

6.4 Reports to State Departments of Education and to schools

To better motivate participation in the assessment, schools that so request should be sent a computerized report of score distributions for their school expressed in percentiles of the State and national distribution. These reports can be similar to those of State assessment programs reporting by school, and copies of the report can be sent to the corresponding State departments of education.

6.5 Data for linking the National Assessment scales to those of State Assessment programs

For those States that conduct an all-student assessment census at the same grade levels as the State-reporting assessment, it is possible to link the scales of the State assessment program to those of the national scales by matching the NAEP data file for the State with the individual student file of the State assessment program. Matching is done in the following way. When NAEP tests in a school it leaves with the school a roster of NAEP identification numbers for students who are tested. Although NAEP does not retain a copy of these rosters, and thus does not know the local identity of the students, the NAEP ID numbers appear on the case records in the NAEP internal files. With the cooperation of the schools, the State assessment program can obtain the school rosters and match their student case record numbers with the NAEP ID numbers. Once these matches have been made, it is a fairly simple matter to develop equations predicting scores on the national proficiency scales, from one or more students' scores in relevant subject matter areas, from the tests administered by the State assessment program. By means of these equations the much more detailed State-level results provided by the State assessment program can be expressed in the units of the national assessment scales. This procedure is an alternative to augmentation of the State sample in order to provide within State comparisons. (See also 7.4.)

6.6 Secondary user files

Many of the policy-relevant analyses of national assessment data must be made by secondary users in the universities and research institutes from data files provided for this purpose. Because such analyses are often difficult and time consuming, it is essential that the user files be available in a timely manner, preferably by the January following the assessment year. These files should contain student-level item responses, student- and school-level scale scores and all questionnaire responses. They would not identify students or schools but would be organized by identified States.

7. Technical Support

7.1 Assessment-instrument maintenance

7.1.1 Free-response and multiple-choice tests

Provided the item content is not radically altered, item response theoretic methods make it possible to replace a number of items in an established scale without losing comparability with previously computed scale scores. The use of these methods in a continuing assessment program is necessary as items become obsolete and must be replaced and are released to the public for illustrative purposes. As a matter of policy, perhaps 10 percent of the free response and multiple choice items in each subject matter area should be retired after each assessment. In order to provide a pool of pretested replacement items for this purpose, a few so-called "variant" items should be added at the end a number of forms of the pretest booklet as described in Section 3.3. In addition, statistical procedures for examining the parameters of retained items, especially analyses for predicting drift of item locations on the assessment scales should be routinely carried out after each assessment. Items that are changing excessively can then be replaced, or their parameter values adjusted to account for drift. (See Bock, Muraki, and Pfiffenberger, 1988.)

7.1.2 The writing test

The writing exercises present special problems for maintenance because both the effects of exposure of the writing prompts and changes in the standards of reading must be accounted for. Moreover, a special type of item-response model applicable to graded ratings of the written essays must be employed so that new prompts can be introduced, and old prompts retired, without changing the interpretation of the writing proficiency scale. Procedures developed for the California Direct Writing Assessment could be used for this purpose. Reading standards are checked and accounted for by including in the readings for the current assessment a small proportion of essays from previous assessments. Ideally, the reader should not know whether they are reading a new or old paper. The new scores on the old papers can then be compared with previous scores in order to estimate and correct for any changes of performance of the reading teams. Once these checks and corrections have been carried out, average scores for old prompts in the current assessment can be examined in relation to those of the new prompts and values from previous assessments in order to detect possible effects of exposure of the prompts and coaching.

7.1.3 Questionnaire

Questionnaire items will need to reflect to some extent changing theoretical interests among educational researchers. Revision of the questionnaire at 4-year intervals should be sufficient for this purpose.

7.2 Performance-test development

The recommendation is that a performance testing unit, with continuing funding, will function collaterally with the main assessment program, developing performance tests in the various subject matter areas and administering them in special individual testing sessions in participating schools, following the main assessment. As suggested in Part I, results of this type of testing can play an important role in establishing the practical meaning and validity of the assessment scales.

7.3 Scale validity and item writing techniques

The procedures and item content of the assessment should keep abreast of new developments in the fields of curriculum and cognitive science that have implication for construct validity and item writing techniques. The budget of the assessment operation should include funds to support continuing studies in these areas, possibly through grants to research centers and universities.

7.4 Linking of national assessment scales to those of assessment programs in the participating States

As discussed in Part I, detailed information about subpopulation or program effects within States could be obtained by augmenting the State reporting assessment. If the State already has a "census" assessment in which all students are tested, however, a more economical approach would be that of linking the scales from the State assessment program with those of the national program, so that within State results could be compared with the Nation or with those of other States. Relating the State scales to the national scales is straightforward if records for students within the States who have participated in both the national and the State assessment can be matched as described in Section 6.5 of Parts II and III. From the matched files, standard statistical procedures for estimating linear relationships can be applied in order to predict scores on the national scales from those of the local State assessment program. At the aggregate level of districts, programs, or subpopulations, these predictions can be quite accurate. They would enable the State to compare specific within-State data to national performance, or to that of other States who have carried out similar linking procedures.

7.5 Methodological development

Because changing concerns in education and new technology will require the assessment procedures and materials to evolve over time, it is important that funding be available for continuing methodological development. Especially important are orderly methods for making procedural changes within the assessment without losing continuity in the assessment scales. The general principle should be that any proposed innovation be developed in prototype and tested concurrently with the existing procedures. These tests are perhaps best carried out by over-sampling students in a sample of larger schools throughout the Nation and assigning, within those schools, half of the students to the old procedure

and half to the new procedure. The data from these trials, paired by school, would then provide a sensitive test of differences in the procedure and would estimate the effects required to equate the new and old reporting scales.

A high priority task for the development team should also be the exploitation of new computer technology to permit more flexible response modes than is possible with the simple mark-sensing equipment now employed to score multiple-choice items.

References

Bock, R. D., & Mislevy, R. J. (1988). Comprehensive educational assessment for the States: the duplex design. In Educational Evaluation and Policy Analysis (in press).

Bock, R. D., Muraki, E., & Pfiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. In Journal of Educational Measurement (in press).

Measurement Objectives for State Assessments by NAEP

John T. Guthrie
Susan R. Hutchinson
University of Maryland

Policy Context of State-NAEP Assessments

According to the Federal legislation for State-by-State NAEP Assessments and perspectives of the National Governors Association and the Council of Chief State School Officers, the State-NAEP Assessments will take place in a policy context. These sources suggest that the purpose of State-NAEP Assessments is to obtain measures of educational achievement in order to inform and focus discussion and debate regarding educational policy for the improvement of schooling

The State-NAEP Assessments are intended to provide comparisons of educational achievement across States. They are also expected to provide comparisons across time within a single State. Beyond these normative data, descriptions of achievement in absolute terms will be obtained. Statistics will describe how well students are accomplishing educational objectives in reading, writing, math, science, geography, history, and literature.

Assessment in a policy context implies that the results will be used for quality control of the educational system. When assessments are used for discussion about policies for school improvement, the primary issue is usually identifying and allocating educational resources. If educational administrators and political leaders judge achievement to be unsatisfactory or inadequate, increased funding, personnel, training, or other resources are usually devoted to the identified problem. The problem area may consist of a school subject such as reading or science or a subpopulation such as low income students that will be targeted for special programs. Resources may be reallocated toward development of specialists, improvement of learning materials, changes in scheduling, increases in classroom time, inservice teacher training, local educational assessments, school organization (schools within schools), parent involvement programs, and others (Cohen, 1988).

Policy shifts are fundamentally directed to improving student learning. Student achievement is the most appropriate goal of policy initiatives. In the State context, assessments can have substantial impact on instructional goals and curriculum objectives. As Connecticut's Commissioner of Education, Tirrozi (1985) states, tests become goals for instruction when schools are held accountable for performance on them. Further, as Railsback (1987) claims, the measurement objectives of NAEP are attractive to practitioners because they are developed from a broad base by reputable leaders. As a consequence of these forces, it is highly likely

that the measurement objectives of State-NAEP will impact the educational objectives of States, districts and schools.

Interactive Models of Knowledge and Cognition

If State-NAEP takes a leading role in directing student learning, the measurement objectives of State-NAEP should reflect the contemporary research on student learning and cognition. If assessment is to influence the nature of teaching, then what is known about learning and cognition should influence the design of the assessment.

Seven years ago, Glaser (1981) predicted that cognitive science would soon be in a position to inform psychometrics. By describing the nature of expertise in an area such as text comprehension and math problem solving, cognitive psychologists could describe the more and less important components of processing and the relationships among them. This information could lead to specifications for more accurate and valid tests. Since Glaser's forecast, substantial progress has been made in reading, math, writing, problem solving and knowledge acquisition.

A fundamental quality of the scientific models of proficiency in these domains is interactivity. That is, cognitive components of processing in reading, math, problem solving and knowledge acquisition are interdependent and mutually reliant. Examples will be given briefly in reading and math.

Several chapters in the Handbook of Research on Reading illustrate the interactivity among the following:

1. Background Knowledge
2. Metacognitive Processes
3. Paragraph Comprehension
4. Inferencing
5. Sentence Comprehension
6. Vocabulary Knowledge

The naive view of these processes is that the higher-level operations such as gaining knowledge from reading or performing metacognitive operations, are dependent on lower level operations such as knowing vocabulary or sentence comprehension. Interactive models, however, specify reciprocal influence. Processing is top down as well as bottom up. The literature is replete with examples and experimental evidence. Knowing the meaning of a word determines comprehension of the paragraph, but also comprehension of the paragraph determines the depth of understanding and learning of the words within it.

The interactive models suggest that components cannot be measured accurately in isolation. One cannot measure the proficiency of higher-order skills without being certain that lower-order skills are attained to a sufficient level; and one cannot determine the mastery of lower-order skills without an estimate of the level of higher-order skills which influence them. As a consequence measurement of these processes must be unified and simultaneous. Separation of processes within reading is theoretically

indefensible.

Mathematical proficiency has been construed as four classes of knowledge by Leinhardt (1988). These are depicted as interactive components of mathematical understanding.

1. Principled conceptual knowledge
2. Computational knowledge
3. Concrete knowledge
4. Intuitive knowledge.

Without extensive definitions and elaborations, principled knowledge consists of the mathematical propositions that justify or constrain procedures. Computational knowledge is the procedural knowledge of the algorithms and operations such as addition and multiplication. Concrete knowledge is understanding of nonalgorithmic systems such as pie drawings to represent fractions. Intuitive knowledge is real-life understanding of quantitative circumstances, such as how to avoid being cheated in a game that involves numbers. The acquisition of these processes over time during instruction illustrates that they are interdependent. One cannot be "mastered" before others are raised to sufficient, supportive levels. Full maturity of any one process relies on proficiency of the others. As a consequence, assessment of these processes must be simultaneous and dynamic. An assessment model that assumes independence among processes is not likely to provide a valid representation of mathematical understanding (Carpenter & Peterson, 1988).

Validity of State-NAEP Tests

The information that is inherent in the interactive models of reading or math has implications for State-NAEP assessments. The primary implication is that the measurement objectives should represent the interactive models of the domains being tested. In particular, construct validity should be maximized.

The State-NAEP should be oriented to the goals of describing student competence and assessing program effectiveness. It should not attempt to provide diagnostic information that will be used to prescribe differentiated skill instruction for individual students. For these purposes Cross and Paris (1987) recommend that construct validity is an essential property. In their view tests such as State-NAEP should measure the content domains and cognitive processes broadly and deeply as they are understood by experts. For example, a measure of problem solving in math that has construct validity will fully tap the cognitive operations of problem solving as they are defined in the research literature and broadly sample mathematics as it is mapped by content experts. Mere prediction of these competencies via highly correlated tests will not be sufficient.

Test developers often attempt to maximize construct validity by following the scheme described by Haertel (1985). Haertel recommends a three dimensional framework consisting of: 1) Contents. This may include subtopics in a field such as history or subdomains in a field such as

mathematics. 2) Processes. This refers to the cognitive operations such as recall, computation, inference, or problem solving, etc. that is expected for a specified content. 3) Contexts. This refers to whether the task is typically performed at home or school, whether it is speeded or unspeeded, whether it is aided with feedback, etc. A test that is representative of these three dimensions is regarded as having construct validity. Another use of the term construct validity refers to the extent to which the factor structure (of items on a test or subtests on an assessment) is consistent across age groups, ethnic groups or other samples of the population. This latter use of the term literally refers to consistency of the measure of the construct across subgroups and will not be emphasized in this discussion. In this paper the term construct validity emphasizes the extent to which the measure reflects current theoretical understanding about content and process of the achievement area being measured.

Separating content and process. Scores in a content area such as history should not be influenced by whether students are good or poor readers. Likewise, an assessment of a process such as reading or critical thinking should not be contaminated by variations in content knowledge required to perform the assessment exercises. This point seems exceptionally elementary. It has not been accounted for, however, in previous NAEP assessments. The history and literature tests for 17-year-olds required high levels of reading (Ravich & Finn, 1987). While the authors claimed that students seem to lack knowledge of history, the students may lack sufficient reading skill to exhibit their knowledge on the test. There was no control for readability of the items. It is probable that the bottom quartile of students according to a reading measure were unable to read well enough to perform the items and that they likely depressed the scores of the population. At the same time, reading measures in NAEP have not consistently used highly familiar content in the exercises. Consequently, students who lacked broad general knowledge may have been disadvantaged on the reading assessments of the past.

Separating content and process does not occur automatically using a content and process matrix. Summing performance on items in a content category across process categories will result in the contamination of the two as has prevailed in the past. To distinguish content and process, exercises must be individually constructed to meet the simple requirement that there are low process demands for content measures and low content demands for process measures. Scores on an assessment of history knowledge should not be contaminated by variance in processes such as reading inferencing, critical analysis or metacognition. Likewise, scores on an assessment of critical thinking or problem solving should not be confounded with content that is known by some students and unknown to others. To optimize construct validity of State-NAEP assessments, the measures of process and content should be more distinct from each other than they were in previous NAEP assessments.

Content Validity

When assessments are used for State comparisons, the dissimilarity of curricula in subjects such as history, literature, science, math, geography or civics raises the problem of content validity. Universal comparability, or curriculum equity, requires the "least common denominator" approach in which objectives common to all States are tested. This criterion leads to a narrow and possibly regressive assessment and fails to spur schools toward expansion and extension of education. The second basis for content validity of an assessment is expert judgment. A select group of experts highly reputed in each content domain, for example historians, write the specifications for topics and subtopics in that domain. This criterion may lead to a high level of curriculum-test alignment in a few States but lower levels will be seen in other States. Thus comparability of the assessment may be compromised. If substantial agreement among participating States is obtained for objectives that are proposed by experts and professionals, however, the assessment will represent desirable content. It will be valid by the standard of expectations and ideals of professionals, though it will not be strictly equitable for existing curricula.

A third approach to content validity is based on a composite. Two groups of objectives are identified and combined. The first is the set that is contained in the vast majority of teaching programs, i.e., the least common denominator. The second is the extension that is needed to satisfy the expert specifications. Both sets could be administered; separate scores and a total score could be obtained. State comparisons could be provided for the core, for the extension, and for the total. This provides assessment of student knowledge on what is taught (core), what should be taught (total), and what should be added to the existing curriculum (extension). Each State will be comparable in terms of the curriculum using the core assessment, and comparable in terms of professional expectations using the total assessment. The composite permits the strength of both approaches to be incorporated.

Construct Validity

Interdependences of processes. To establish construct validity for measures of process the implications of models of cognitive processing should be considered. These models suggest, for example, that competence in text comprehension requires word knowledge, sentence comprehension, intertext inferencing, monitoring for misunderstanding, and the use of previous knowledge to understand new ideas. Experimental evidence shows that proficiency in one process enhances performance in other processes (Pearson, 1984). A psychometric prediction from these theoretical models is that processes of reading comprehension will be correlated. Zwick (1987) tested this prediction by assessing the dimensionality of NAEP reading data. She used the 1983-84 NAEP data for 9-, 13-, and 17-year-olds, including about 26,000 students per group. About 100 multiple choice items per grade were analyzed. Account was taken of the BIB design. According to two procedures, principal components analysis and full-information factor analysis, the test appeared to be unidimensional. One factor solution was optimal in both cases. Although the test contained

items measuring literal comprehension, inferential comprehension and interpretation, the items tended to form a single factor. This unidimensionality is predictable from the cognitive models of reading.

Models of mathematical thinking contain interactive cognitive processes. From these models, it can be predicted that NAEP math assessements will reveal interdependences among process. Suchner (1988) reanalyzed the 1985-86 assessment for 13- and 17-year-old students, with over 25,000 per group. Process objectives were. 1) skill in mathematics, 2) knowledge of mathematical concepts, 3) routine application, 4) understanding/comprehension, 5) problem solving/reasoning. These were assessed across seven content areas such as discrete mathematics, measurement, geometry, numbers, and operations. Principal component analyses showed that the first component accounted for 72 percent of the variance. The other components had eigenvalues lower than 1.0 and accounted for small proportions of variance. As the author concludes, there is a high degree of convergent validity within the math domain. In an investigation of math achievement in elementary school children, Klein (1985) found that for both a State and national sample of fourth- and eighth-grade students, three mathematics subtests, consisting of concepts, problem solving and computation, clustered together on a single oblique factor. The theoretical and psychometric dependencies among processes in math seem to be as substantial as they are in reading.

Studies by Wikoff (1978), Roberge and Flexer (1981) and Hanna and Lei (1985) also reported factor analyses for fourth through twelfth-grade students in which mathematics subscales loaded on single factors. Roberge and Flexer (1981) found that three math subtests, computation, concepts and problem solving, included in the Metropolitan Achievement Tests, had high loadings on a general intelligence factor Hanna and Lei (1985) showed that the relationship between two subtests of the Canadian Tests of Basic Skills, math concepts and math problems, was consistent in grades four through six irrespective of differences in curricula. In sum, factor analytic studies confirm the expectation from cognitive research that math processes are highly associated and are indistinguishable according to psychometric criteria.

The scaling, scoring and reporting of separate processes within reading and math does not appear to be supported by cognitive process models or psychometric properties of the scales. In a State-NAEP assessment of math, the construction of scales such as computation, procedures, estimation, or problem solving, is not easily justified. Divisions such as these are likely to yield reductions in construct validity. Although an assessment in a given content area, such as math or reading should require the fundamental processes, the interdependencies among them are so high that dividing them into subscales and subreports is likely to be misleading.

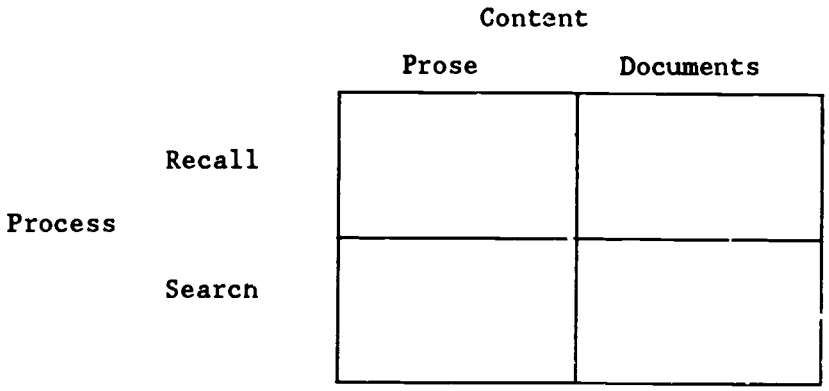
An alternative to multiple scales for subprocesses in math or reading is one scale for each subject matter, such as math, at each grade level. For purposes of informing the educational system, however, a single scale per subject is not sufficient. Math learning, or science achievement, or reading proficiency are vastly too complex to be represented by a single

core. Professional educators at the school, district and State level require a richer display of information than a "g" score for each learning area. The ideal condition for State-NAEP is to develop and report distinct constructs.

Developing Distinct Constructs

It is likely that more than one distinct construct exists within each area of reading, math, and writing. There will be at least two intersections of a content-process matrix that are relatively unique theoretically and psychometrically. Such constructs will be unique theoretically if they are influenced by variables that do not influence other constructs or if they contain processes not contained in other constructs. They will be unique psychometrically if they are not highly correlated with each other.

A simplified content-process matrix in reading includes the following:



The intersections of prose recall and document search (e.g., upper left, and lower right) each consists of a form of written language that is processed with a typical procedure. Such intersections are likely to be distinct constructs. These two constructs in reading were isolated in the NAEP adult literacy report. These consisted of: 1) comprehension of prose, and 2) search of documents. Kirsch and Jungeblut (1986) reported that these factors were correlated at about .5. In addition, Guthrie and Kirsch (1987) found that document search and prose comprehension were distinguishable in a factor analysis. In these studies document search requires the student to locate details in tables, charts or prose; whereas reading comprehension requires the use of strategies for learning and remembering knowledge from expository or narrative prose. Both the cognitive models and the psychometric properties of measures of document search and reading comprehension have distinctive features and characteristics. These two constructs appear to warrant independent assessment and reporting in State-NAEP reading measures.

Distinctive constructs within writing (Stein, 1986) and math (Leinhardt, 1988) have been proposed conceptually, and examined with

167

experimental or observational studies. However, they have not been widely used in assessment. In math a simplified content-process matrix may include:

	Numbers and Measures	Algebra/ Geometry
Concrete Knowledge/ Computation	1	
Principled Knowledge/ Problem Solving		2

Measures of 1) computations with numbers and measures and 2) problem solving in algebra/geometry are likely to be distinctive. They are plausible candidates for relatively unique psychometric constructs. These measures are consistent with the cognitive perspective that a complex set of processes exists within each construct, i.e., 1 and 2, that should not be artificially segregated in testing, scaling or reporting. Furthermore, they provide an increase in detail compared to single scores. Such an approach to construct development, scaling and reporting acknowledges the complexity of competence in math, while retaining the cognitive and psychometric criteria for high caliber measurement.

An implication of the cognitive models for development of distinctive constructs is that the measurement of a given process should be contextualized within measurement of other processes. For example, the strategies of reading should be measured interactively with vocabulary and prior knowledge in a long text. As the first segment is being read, questions should be asked about use of prior knowledge, detection of inconsistencies within text and predictions about meanings that will occur in later text. Following another reading segment, necessary inferences, ability to summarize previous material, and abstraction of theme could be measured. In a final section metacognitive skills such as knowing when to check one's understanding and how to detect failures of comprehension should be tested. In such an assessment, items are not independent. This is a positive feature rather than a negative one, however, because the processes being measured are not independent.

Policy Implications

Because State-NAEP is a policy-driven assessment, the implications of possible results should be considered. States or districts that receive relatively low scores on the assessment are likely to direct more resources to their areas of weakness. School improvement efforts, furthermore, will be guided by the constructs, such as mathematical competence or knowledge of history that are defined in the State-NAEP assessment. In States with

accountability systems, the tests represent complex tasks that may become the focus of reform in educational settings. Revisions of curricula, instruction, school organization and leadership may be directed to educational goals that are defined by measurement objectives. If these educational goals are significant rather than trivial, and based on research rather than other factors, they will be productive. Developing measures for State-NAEP that optimize construct validity as well as content validity, will contribute significantly to policy judgements that are educationally effective.

References

Bock, R. D., & Misley, R. J. (in press). Comprehensive educational assessment for the states: The duplex design. Educational Evaluation and Policy Analysis.

Carpenter, T. P., & Peterson, P. E. (Eds.) (1988). Educational Psychologist, 23(2).

Cohen, M. (1988). Designing state assessment systems. Phi Delta Kappan, 69(8), 583-589.

Cross, D. R., & Paris, S. G. (1987). Assessment of reading comprehension: Matching test purpose and test properties. Educational Psychologist, 22(3 & 4), 313-332.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 36(9), 923-936.

Guthrie, J. T. (1987). Indicators of reading education. Center for Policy Research in Education Monograph. The State University of New Jersey, Rutgers, The Rand Corporation. New Brunswick, NJ.

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. Journal of Educational Psychology, 79(3) 220-227.

Haertel, E. (1985). Construct validity and criterion-references testing. Review of Educational Research. 55(1), 23-46.

Hanna, G., & Lei, H. (1985). A longitudinal analysis using the LISREL-model with structured means. Journal of Educational Statistics, 10(2), 161-169.

Kirsch, I. S., & Jungeblut, S. (1986). Literacy: Profiles of America's Young Adults. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Klein, A. E. (1981). Redundancy in the Iowa tests of basic skills. Educational and Psychological Measurement, 41, 537-544.

Leinhardt, G. (1988). Getting to know: Tracing students' mathematical knowledge from intuition to competence. Educational Psychologist, 23(2), 119-144.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21(3), 215-327.

Pearson, P. D. et al. (1984). Handbook of Reading Research. NY: Longman.

Pearson, P. D., & Gallagher, M. C. (1983). The instruction of reading comprehension. Contemporary Educational Psychology, 8, 293-316.

Pressley, M. (1986). The relevance of the good strategy user model to the teaching of mathematics. Educational Psychologist, 21(1 & 2), 139-161.

Railsback, C. E. (1987). Using the national assessment to improve student achievement. The Reading Teacher, 66-73.

Ravich, D., & Finn, C. E., Jr. (1987). What do our 17-year-olds know: A report on the first national assessment of history and literature. NY: Harper and Row.

Roberge, J. J., & Flexer, B. K. (1981). Re-examination of the covariation of field independence, intelligence and achievement. British Journal of Educational Psychology, 51, 235-236.

Stein, N. (1986). Knowledge and process in the acquisition of writing skills. In Ernst Z. Rothkopf (Ed.), Review of Research in Education. 13, 225-258.

Suchner, R. W. (1988). Modeling NAEP items by cognitive process. Paper presented to the Annual Meetings of the American Educational Research Association, New Orleans, LA.

Tirozzi, G.N., Baron, J. B., Forgione, P., & Rindone, D. (1985). How testing is changing education in Connecticut. Educational Measurement: Issues and Practices, 4(2), 12-16.

Wikoff, R. J. (1978). Correlational and factor analysis of the Peabody Individual Achievement Test and the WISC-R. Journal of Consulting and Clinical Psychology, 45(2), 322-325.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. Journal of Educational Measurement, 24(4), 293-305.

Collecting and Profiling School/Instructional Variables as Part of the State-NAEP Results Reporting: Some Technical and Policy Issues

Joan Boykoff Baron
and
Pascal D. Forgione, Jr.
Connecticut State Department of Education

The primary purpose of the NAEP program has been to provide data about what our Nation's students know and can do. A State-NAEP program would have the same primary goal. However, to the extent that national, State and local district policymakers attempt to use these achievement indicators to formulate educational policy, they will be concerned with the related question, "How can we improve our students' achievement?" Basic and applied educational research over the past several decades has yielded considerable insight into variables which influence the teaching-learning process. If National- and State-NAEP can incorporate these insights into a minimally intrusive data-collection strategy, then policymakers will have some valuable guidance in developing effective policy.

Throughout this paper, the primary data set used to generate our conclusions and examples is from the Connecticut Assessment of Educational Progress (CAEP) Program. For each assessment conducted since 1971, we have created extensive student, teacher, and principal questionnaires. Some of our questions originally came from NAEP and studies of the International Association for the Evaluation of Educational Achievement (IEA) and our findings about the importance of certain variables have been consistent with the findings reported by those large-scale studies. In addition, our findings are consistent with those reported by our neighboring States of Massachusetts and Maine who have used many of the same items on questionnaires administered with their statewide assessments. On each assessment, we were guided by advisory committees who had many ideas about what variables were important to measure. However, because we did not have available the reductionistic guidelines provided in this paper, we often collected data on numerous dimensions that had no relationship to students' achievement and/or had no implications for educational policy. In fact, more than half of our questions, although they provided "interesting" data, had no direct utility for either understanding or improving education.

Because of the impending scope of State-NAEP and the testing time it will require, we recommend a judicious selection of questionnaire items. If properly chosen, their potential impact is enormous. If carelessly

The authors gratefully acknowledge the suggestions made by the following people on earlier drafts of this paper: Joan Allen, Beverly Anderson, Elizabeth Badger, Leigh Burstein, Neil Carey, Pat Cox, Allan Hartman, Hannah Kruglanski, Kristine Mika, Cynthia Prince, Douglas A. Rindone, Amy Shively, and Grant Wiggins.

chosen, not only is their impact minimal but the intrusion into students' learning time is unforgivable. It is equally inappropriate to burden teachers and principals with lengthy questionnaires unless the data have direct relevance to educational policy and/or practice.

In developing this paper we did not begin with a set of abstract principles about what makes a questionnaire item useful. Rather, we began with a massive array of questionnaire items, only a small fraction of which had provided useful data. It was in the categorization of the useful questions that two major guidelines emerged. (The major portion of this paper provides examples of these guidelines).

- o A variable should either be directly related to achievement or be a highly valued outcome of the educational process, which may not necessarily be related to achievement as measured by the test.

- o Any variable that is related to achievement should meet one of the following three conditions: (a) it belongs to a class of "unalterable" variables that will be used as reporting categories for subpopulations; or (b) it is useful for the establishment of the NAEP test's concurrent validity; or (c) it is both "alterable" by schools and highly valued in its own right.

Any variable which does not meet the above criteria should not be included. This would eliminate hundreds of questions which are not related to achievement or other desired outcomes of schooling and have no direct implications for policy or practice.

The major focus of this paper will be on the "alterable" variables, which do relate to achievement with an emphasis on those involving the instructional process and its context. However, before discussing the schooling variables we will scratch the surface of two of the other legitimate uses of questionnaires mentioned above.

Unalterable Variables Used to Report on the Achievement of Subpopulations

Questionnaires are the best way to collect background and demographic data on unalterable variables that can be used to report results by subpopulations. Examples of these unalterable variables include gender, race/ethnicity, socioeconomic status, and type of community. They are considered "unalterable" because schools cannot alter a student's group membership. However, whereas these variables cannot be altered, the relationship between the variable and achievement is considered to be alterable. That is, in the future, we hope to witness the reduction and ultimately the elimination of the relationships between these variables and student achievement. (See Forgione and Baron, 1987.) In fact, most of our Nation's educational policy rests on this assumption. Although it is not the primary focus of this paper, we recommend that to the extent possible data on these unalterable variables be collected from the students as opposed to from census data tapes. Using census data makes disaggregation within schools impossible and would treat an entire school population as interchangeable. (This is a complex issue because when there are critical

concentrations of groups of students, the school climate is altered; See Burstein (1980) for a treatment of this concept.)

Establishing the Test's Concurrent Validity

A second group of variables may be useful for the establishment of a test's concurrent validity. In Connecticut, two questions about students' self-reported school achievement have sufficed for this purpose. They are:

- What grades do you usually get in school? and
- What grades do you get in science? (or whatever specific area is being tested).

Both of these questions have demonstrated very strong positive relationships to students' achievement in every test we have administered. That is, students who perform well in school perform well on our tests and those who perform poorly in school perform poorly on our tests. This finding gives us greater confidence that our tests and school grades are measuring similar knowledge, skills, and motivational networks.

Alterable Variables

Included in the potential universe of alterable variables are many of the indicators catalogued by Oakes (1987) under Access to Knowledge, Press for Achievement, and Professional Conditions for Teaching. Within each category, Oakes further subdivides the variables into three types: Resources, Structures, and Culture. We are summarizing the Oakes classifications because they represent a fairly comprehensive set of important variables that have been linked to achievement. (However, we do not subscribe to the view that they are all important to assess in either National- or State-NAEP for reasons stated at the end of these sections.) In the grouping called Access to Knowledge, Oakes (pp. 31-32) is concerned with the issue of whether students of all abilities have sufficient opportunities to learn.

The Resource variables include the availability of sufficient instructional materials, laboratories, computers, and equipment; teachers' qualifications and experience for the courses they teach, and the availability of discretionary funds for supplies, materials, trips, speakers, etc. Under Structure, Oakes includes instructional time in days per year and hours per day, the emphasis the school places on different curriculum areas as measured by course offerings and staffing patterns, the procedures schools use to assign students of different abilities to classes and the types of assignments they receive within classes, as well as the academic enrichment and supports available to students. In the Culture section, Oakes is concerned with opportunities for staff development, parent involvement and staff perceptions about the importance of learning for all students.

In Press for Achievement, Oakes (pp. 33-34) is concerned with how schools organize their staff, time, curriculum, and materials to support the belief that all students can learn. The Structures include student

participation in long-term projects, papers and research activities, opportunities for school-wide recognition of accomplishments, graduation requirements, and student participation in challenging study as measured by enrollment in challenging courses, and average course completion rates. The Culture variables are more diverse and include graduation and attendance rates as well as student attitudes toward achievement and staff perceptions about the importance they and their school place on student achievement.

In Professional Conditions for Teaching, Oakes (pp. 35-36) is concerned with how schools provide teachers with the supports regarded as important in order to be successful on the two categories described above. The Resource variables include teacher salaries, pupil load, class size, funding for school-based staff-development activities, and clerical support available for teachers' noninstructional tasks. The Structures include the amount of teacher-time scheduled for teaching, non-teaching work, school-wide staff-development activities, and special teacher-developed projects (e.g., curriculum development, instructional improvement, collaborative research, etc.). The Culture variables include a set of staff perceptions related to the school's goals and the nature and level of staff involvement in curriculum and instruction.

We do not advocate using all of these indicators on either National- or State-NAEP for at least two reasons: First, the data burden would be unacceptable. Second, some of these variables are difficult to operationalize on questionnaires and may best be obtained from visiting a school and observing its climate. In the remainder of this paper we will provide examples of how some of these indicators as well as others not on this list have proven to be effective on questionnaires used in Connecticut and elsewhere. The choice of which variables to include will require establishing a set of priorities by educational policymakers. Hopefully, the guidance provided in this paper will assist them in their task.

Many alterable variables are broad in their scope of influence and affect all school learning. Hence, the same questions can be asked on all NAEP assessments. Others, on topics such as: "Opportunity to learn" (i.e., exposure to the content of the test) and those "expert" behaviors exhibited by effective readers, writers, mathematicians, scientists, etc. are closely tied to specific subject domains. By extension, this second group of variables also includes those dimensions of "effective instruction" that have been demonstrated in research and evaluation studies to enhance student achievement. In most cases these practices attempt to foster in less effective students the behaviors engaged in by the more effective (or "expert") students. This is the category that promises the most reward, and like most valuable commodities is the "hardest to come by." This will be examined below in the section entitled "The Difficulty of Capturing the Instructional Process on Questionnaires." In the remainder of the paper we will share some formats and data that have been useful for us. This is done to demonstrate the value of questionnaire items that get at the "heart of the educational process."

Reporting Alterable Variables on Which "More Is Better"

If data is reported publicly on a certain cluster of alterable variables, implicit in that reporting is the clear message that "more is better." Therefore, for any alterable variable on which data is reported, it should be desirable for schools to increase the numbers of students who indicate the presence of that variable. In selecting those variables, it is essential to ask two questions:

- Would educators value an increase in the number of students reporting on the presence of that variable? and
- Is the indicator corruptible? i.e, Is is possible to report an increase on the variable with no concomitant increase in achievement?

Murnane (1987) noted that it is possible for schools to corrupt the relationship between certain variables and achievement. His example concerned the number of math courses taken by students. He argued that students could take a greater number of "watered down" mathematics courses with no resultant increase in their mathematics achievement. This concern finds support in the Underachieving Curriculum (McKnight, et al, 1987) which noted that United States mathematics achievement was low despite a substantial amount of class-time spent in mathematics instruction in this country. This corruptibility is probably not intentional, but rather is a reflection of a misunderstanding of what is important.

It is not the sheer number of courses or hours spent on mathematics. but rather the strength of the curriculum dispensed during those hours. McKnight et al. stress the importance of "curriculum as the distributor of opportunity to learn" (P. 85). Corruptibility can be diminished by indicators that are more specifically defined and include a qualitative as well as a quantitative dimension. It is also important to look for unintended side effects of changing or adding an indicator. This can be accomplished by broadening the domain of variables monitored. Koretz (1988) recently described the effects of a high school's increasing its foreign language requirements from 1 to 2 years. The major result was that more students took 2 years of a language. But, the offerings of advanced courses were severely decreased because there weren't any teachers available to teach those courses. This demonstrates the importance of broadening the indicators to include some unintended side effects. Therefore, those devising the NAEP questionnaires should attempt to think through, in advance, how each variable might be corrupted and impact upon other parts of the educational enterprise. These deliberations would be used to more carefully define the variable of immediate interest as well as those potentially related to other relevant aspects of the system.

Highly Valued Activities Which Are Not Related to Achievement: Examples from Science, Writing, and Social Studies

The lack of relationship between a variable and students' achievement may exist for several reasons. Three of these are described below.

Insensitive Measures of Achievement. Sometimes, the lack of relationship between a variable and achievement may be due to the insensitivity of the total test scores. For example, on our fourth grade science test, the use of scientific apparati was not related to the students' total science scores (See Exhibit 9). On the 8th grade test, however, we demonstrated that prior use of a triple-beam balance in class was directly related to the specific performance tasks on our tests which required using that piece of equipment (See Exhibit 10). (This particular item tests for near transfer and does not require much in the way of far transfer; a comprehensive assessment would attempt to include both.) Our concern is that important instructional variables not be disregarded when they have a strong impact on a narrow range of achievement that can be washed out when only total achievement scores are considered.

Assessing Necessary but Not Sufficient Activities. The presence of some variables may constitute a necessary, but not sufficient, condition for improvement. For example, on our CAEP writing assessment in English Language Arts, we used a series of NAEP questions about how many papers students had written in the previous 6 weeks. There was a strong relationship between number of papers written and the students' writing scores. Therefore, it might be fruitful to track the number of papers students write, because without ample writing opportunities (a necessary condition), it will be difficult to improve students' writing. Yet, if nothing of value happens to foster better writing, it is unlikely that students' writing achievement will increase as much as it would if students' practice were coupled with the employment of effective writing habits.

Some Behaviors Valued in Their Own Right. There may be some behaviors that are valued, independent of their relationship with achievement. Examples from our 1982-83 CAEP Social Studies test include data on the frequency with which students vote in school elections and their involvement in different kinds of community service activities. Whereas neither of these variables was directly related to school achievement, many social studies teachers were interested in the patterns of student participation.

Measuring Students' Attitudes: As Important in Their Own Right Examples from Science

Students' attitudes are often related to their achievement, yet one doesn't know the usual direction(s). Occasionally, there may be some attitudinal dimensions related to a subject area that policymakers want to monitor because they are important in their own right. Some examples from science are two statements with which students were asked to agree or disagree:

- "Careers in science are more appropriate for men than for women," and
- "My knowledge of science will be of little value to me in my day-to-day life."

While it is encouraging that only 14 percent of our eighth graders and 11 percent of our eleventh graders agreed with the sexist statement, it was highly discouraging that 42 percent of the eight graders and 41 percent of the eleventh graders did not see the value of knowing science in their everyday lives (See Exhibit 1). It is interesting to note that students who held the desirable attitudes did better on the science test. Yet, even if there had been no relationship, these attitudes might be considered by science experts as important to monitor.

The Difficulty of Capturing the Instructional Process with Questionnaires: Examples from Social Studies, Science and Writing

In Connecticut, we have often tried with little success to capture the value of different types of broad instructional activities on our questionnaires. On several successive assessments in Social Studies, Science and English Language Arts we asked students questions related to the frequency with which they had lectures, discussions, field trips, etc. in their classes. The major conclusion time and time again was that the delivery system did not substantially affect achievement (see Exhibit 2). For most activities, moderate amounts were associated with the highest achievement scores. Therefore, based on our finding, we would not recommended including a litany of such activities on National-NAEP or State-NAEP.

On those few occasions where the delivery system was related to student achievement it was because there was a link between the instructional process and the behaviors that accompany high achieving students. For example, once we knew what good writers do we could begin to understand which specific instructional strategies would enhance good writing. The best predictors of students' writing scores were those items that asked students whether they revised the content of their writing (e.g., added and deleted ideas, moved sentences around, etc.). See Exhibit 3. Therefore, we would advocate developing questions designed to find out whether students actually engage in the practices of "experts" in that domain.

Considerably less predictive of students' writing achievement were questions about the quality and quantity of their teachers' feedback. Not surprisingly, teachers give more feedback to lower-achieving students. Somewhat predictive of students' achievement were behaviors of teachers which encouraged students to use the revisional process (e.g., how often their teachers asked them "to make notes before you write, write the paper more than once before it is graded, reread your writing to yourself, and read your writing to someone else"). Worth noting is that these relationships were much stronger in grade 8 than in grade 11. (However, these relationships between teachers' instructional procedures and students' achievement were still weaker than those asking the students directly about their own writing habits). See Exhibit 4.

It appears that the aspects of the instructional process which are most strongly related to students' achievement are those aspects which directly enhance the development of those skills and strategies that are engaged in by "expert" students in that subject area. Another way to view this is to

try to measure the amount of "time on task" on those classroom strategies which foster the behaviors that are engaged in by the high achievers. Many of these issues will be addressed in the next section.

Assessing Opportunity to Learn

What is becoming clear is that schools need to structure their curricula and their classes in ways which expose students to "the right stuff". The right stuff consists of the total set of experiences--the materials, the techniques, and the apparatus--which foster competence in the area of study. Over time, we have tried a number of different questions designed to find out whether this was occurring for students.

Opportunity to learn has two dimensions--the amount of time spent and the quality of the coverage which occurs during that time. The first of these, the amount of time spent, has several components. One is the number of courses taken--which is a very gross measure of exposure. All things being equal, more courses in an area are better than fewer courses. However, taking a course is a necessary, but not sufficient, condition for learning. It does not include a quality dimension. The same can be said for the amount of time spent studying a subject. (This can be measured by the length of a class, the number of days it meets, etc.) In general, all other things being equal, more time is better than less time. A related variable is the timing of the onset of the exposure to a subject. In some subjects, like foreign language, it seems optimal to begin study at an early age.

The quality of the coverage has to do with whether the time that is spent is spent on "the right stuff." First, does the curriculum include coverage of the major topic areas and the most important skills? That is, does it include exposure to and practice with the tools, apparatus and thought processes required by a particular field? For example, do students use the tools and procedures of the scientist in science class? Do they use the tools and procedures of the mathematician in mathematics class? Do they use the tools and techniques of a graphic artist in graphic arts classes? Do they use the tools and processes of a writer in English and all of their other classes?

Most subtle and hardest to measure is whether the instructional experience is structured so as to improve students' proficiency in the skills required by a subject domain. One of the most promising lines of research occurring today is in the area of "expertise." (Glaser (1987) recently summarized some of the work in this area.) Once we know what experts do, teachers can structure their classroom experiences so as to foster those behaviors in students who have not yet acquired them. Mentoring could be used effectively--having experts (i.e., the teacher and other students) work with novices and think aloud while they work through problems. In this way, students could acquire both the knowledge and the deeper structures of a content area. This view of the instructional process is consistent with that of the teacher as "model and mediator." (See Jones, et al., 1987.) Recent research on the reading process, the writing process, the mathematics process, the science process, and the

process of learning foreign language have the potential to dramatically alter instruction. To the extent that teachers understand what good performers do, they can model, teach and foster those behaviors in their classes.

Extending the School Day . . . into the Home. Schools can do more to make work at home function as an extension of school. The school day is limited. A 50-minute high school class cannot be expected both to teach students new strategies and allow them sufficient time to practice those strategies until they are comfortable with them, internalize them, and apply them appropriately. Homework assignments and long-range independent projects provide ways to meaningfully extend the school experience. Amount of homework functions much like amount of time. All things being equal, more is better. But, homework has a qualitative dimension as well. Meaningful assignments are those which appropriately provide either independent practice or logical extensions of the skills and strategies taught in the classroom.

On CAEP assessments, students are generally asked questions about their homework and sometimes their reading and study habits. Some examples are:

- How much time do you spend doing homework on a typical day?
- Think of the last time you studied for an important test--one that could have had a major effect on your report card grade. About how much time did you spend studying for that test?
- How often do you read the material over a few times when you study for a test?
- How often do you read parts of a story or novel?

Each of these variables exhibited a strong relationship with achievement on the reading test. (See Exhibit 12.) Yet, these questions address the quantitative aspects of homework rather than the qualitative dimension.

Course Taking Patterns: An Example from Science

The Underachieving Curriculum underscored the importance of exposing students to the content covered on the test. At a very gross level, this should be predicted by course-taking patterns. However, Murnane's warnings about corruptibility should give us cause to pause about the long-term value of such a gross measure. On a recent science assessment, we measured "course-taking" patterns at two points in time--before high school and then again, at the end of grade 11. When grade 8 students were asked how many years of high school science they would take, we learned that those students who did not do well on the grade 8 science test were not planning to take as much science in high school as those who did well on the science test. (See Exhibit 5.)

When eleventh grade students reported which courses they are taking or plan to take, certain patterns emerged. The students who scored well on the eleventh-grade test either have taken or plan to take biology, chemistry, and physics. (See Exhibit 6.) A related finding for subpopulations is that more boys than girls enroll in general science,

earth science, chemistry and physics. Only in biology are more girls enrolled than boys. (See Exhibit 7.)

These course-taking patterns considered together with achievement results provide useful information to policymakers. Whereas, 90 percent of our eleventh-grade students take biology in high school, 63 percent take chemistry and 41 percent take physics. The task before school personnel is to find ways to encourage higher enrollment in the physical science courses. However, in doing so, educators should keep not only Murnane's concerns about corruptibility in mind but they should attempt to design courses that are relevant to students' day-to-day life. In this way, more students may report that their knowledge of science may be of value to them in their daily lives.

Is the Content of This Test Covered in Your Classes?: An Example from Drafting, Graphic Arts, and Small Engine

Knowing that our high schools' drafting, graphic arts and small engine programs differ in their inclusion of topic areas and their curricular emphases, both students and teachers were asked whether each domain reported in the results was covered in their classes. This permitted us to report the State results in the context of that coverage (see Exhibit 8). The fact that high proportions of students and teachers report that they do not cover certain domains helps to explain the relatively low achievement in those domains. The Second International Math Study (SIMS) 1987 Opportunity To Learn (OTL) questions have proven to be very successful for measuring content coverage (See McKnight, et al.). Others, including the Ontario Ministry of Education (1988), have replicated the utility of such questions. Currently, Burstein and other researchers are striving to find the best ways to capture OTL data in ways which are minimally burdensome.

Time Spent Studying a Subject: Examples from Foreign Language

Through extensive questioning about the number of hours spent per day, and days spent per week, we learned that the earlier students began studying a language and the more time they spend, the better they perform on the Reading and Listening Tests in the modern foreign languages of French (F), German (G), Italian (I), and Spanish (S). Four specific findings are presented as examples.

- Modern-language students who responded that they started studying foreign language before fifth grade scored higher on the Listening Tests (F,I,S) and on the Reading Tests (I,S) than those starting later; those who responded that they started in high school scored higher on both the Listening and the Reading Tests (F,S).
- Modern-language students who responded that they use their school's language labs for one class period a week scored higher on the Listening Tests (F,I,S).
- Modern-language students who responded that they studied foreign language for 30-45 minutes per day every day for the whole year in grades seven and eight scored higher on the Reading Tests and on the Listening Tests.

- Modern-language students who responded that they studied the target foreign language for more than 15 minutes a day a few times a week before grade seven scored higher on the Reading Tests (F,I,S) and on the Listening Tests (F,I,S).

Classroom Opportunities Which Foster Competence: Examples from Foreign Language

It is not just "being" in class that matters. It is what happens in class that is important. Students who responded that they use the target foreign language in class as opposed to English, perform better on the Reading, Listening, Writing and Speaking Tests.

Modern-language students who responded that they speak mostly the target foreign language or equal amounts of English and the target language in their foreign language class scored higher on the Listening and Reading Tests. Students who responded that they usually speak English in their foreign language classes scored lower on both the Listening and Reading Tests.

- Modern-language students who responded that their teachers usually speak the target foreign language in class scored higher on the Listening and Reading Tests. Students who responded that their teacher usually speaks English in their foreign language classes scored lower on the Listening and Reading Tests.
- Modern-language students who responded that they read mostly in the foreign language scored higher on both the Listening and Reading Tests, (F,G,S). Students who responded that they read mostly English in their foreign language classes scored lower on the Listening Tests (F,G,S) and on the Reading Tests (F,G,I,S).
- Modern-language students who responded that they usually write in the target foreign language in their foreign language classes scored higher on the Reading Tests and on the Listening Tests than those who wrote in English or mostly in English.

Amount of Experience with Apparati and Technology: Examples from Science and Computer Literacy

We asked fourth grade students whether they had used different scientific equipment. Whereas there were virtually no differences in the students' achievement on the total score for the multiple-choice Science test, the percentages of students reporting that they used each piece of equipment have some strong implications for policy. For example, fewer than half the students used a magnifying glass, metric ruler, or thermometer in their science classes. (See Exhibit 9.) However, when we linked students' prior experience with using a piece of scientific apparatus with their specific achievement on the performance test requiring the use of that apparatus there was a clear relationship. For example, grade 8 students who had used a triple-beam balance in the past did better on the performance tasks which required that they used a triple-beam balance. (See Exhibit 10.) In the same exhibit, the data is provided for the amount of experience students report for various scientific apparati. Consistent with their course-taking patterns, most students have used

microscopes and graduated cylinders. Fewer have set up electrical circuits or used triple-beam balances. (It may be worthwhile to disaggregate these data by gender to see whether males and females are receiving equal access to apparatus. In such an analysis, it is essential to control for differential course taking patterns. Therefore, grade 8 data might be the easiest to use for such purposes.)

On our computer literacy test in 1983-84, students in grades 4, 8, and 11 were asked:

- About how often do you presently use a computer in school?
- How many times EVER have you used a school computer to learn or practice computer programming?

More than half of the students in grades 8 and 11 reported that they never used a school computer and those students had lower scores on the computer literacy test. (See Exhibit 11.)

Some Technical Issues in Determining the "Reality of the Classroom"

Wherever possible, NAEP should attempt to link students' achievement to school and classroom variables. This is relatively easy to do if one asks students what happens in their classrooms and schools. It is logistically more difficult if one tries to link the students' achievement in a given classroom to the data reported by the teacher in that classroom or the principal in that school. However, the potential utility in doing so should be worth the added effort.

Several years ago, the State consultant for social studies held a piece of paper up in front of him and asked the first author of this paper to describe what was on it. I proceeded to do so. When I finished, he described what he saw--which of course, was the back of the paper. We were both accurate, but we described different things because we saw different things. That example has stayed with me because it is quite possible that students and teachers describing what is happening in their classrooms may indeed experience different realities. Life seen from the front of the classroom looking back may appear quite different from life seen from the back of the classroom looking forward.

The existence of multiple realities does not mean that we should throw up our hands in despair and not assess the context of the educational process. In fact, we want to argue just the opposite--that in order to understand what is happening in schools, it is important to ask both students and teachers the same questions--with the item stems and options worded identically.

In addition, we believe it is important to ask the same questions over time. For State-NAEP we urge that questionnaire items be piloted extensively before their first administration and once they are put in place, they remain the same over several test administrations.

Data Collected From Students, Teachers, and Principals Related to Oakes' Indicators

For almost 10 years, Connecticut has attempted to measure several motivational and school climate variables. Some of these were originally developed by Brookover et al. (1979) in their research on effective schools. Today, these could be classified under Oakes' three categories: Access to Knowledge, Press for Achievement, and Professional Conditions for Teaching.

Asking Students:

On our 1984-85 Science Assessment students were asked a series of questions which could be classified as Press for Achievement.

- How many students in this school work hard to get good grades on their classroom tests?
- How many students in this school don't do as well as they could in school because they are afraid their friends won't like them as much?
- How often do you come to science class with all the materials you need? (e.g. pens, paper, books, etc.)
- How often do you put a lot of effort into your science homework?
- How much do you agree with the statement, "My teacher cares about how well I do in Science"?

In all cases, students who indicated high Press for Achievement did better on the Science test. (See Exhibit 13.)

Asking Teachers and Principals:

We also asked teachers and principals similar questions that could serve as indicators of Press for Achievement as well as Access to Knowledge and Professional Conditions for Teaching.

Some examples of questions we have asked principals are: (See Appendix A)

- How much difficulty do you have in securing qualified science teachers to fill vacancies? (Grade 11)
- Does your school have a petty cash fund than can be used for science supplies?
- How much does your school annually budget for the purchase of new science equipment (nonconsumable, non-perishable items such as microscopes, scales, etc.--not textbooks)?
- How much does your school annually budget specifically for the purchase of consumable science supplies (materials that must continually be replenished such as chemicals, glassware, batteries, etc.)?
- How many microcomputers does your school have for student use related to science instruction?
- Are your students homogeneously grouped?

Teachers were asked:

- o How available is science equipment (e.g., hands-on materials, glassware, chemicals) for your use in teaching science? (See Appendix C)
- o How well trained are you to teach science at the level you teach? (See Appendix C)
- o In foreign language, teachers were asked how much they use the modern foreign language in class. (See Appendix D.)

Both teachers and principals were presented with a set of factors.

- o "The following set of factors may affect science instruction, and ultimately achievement in your school as a whole. In your opinion, how much of a problem is caused by each of the following?" (See Appendix B.)

  -- A general belief that science is less important than other subjects,
  -- out-of-date teaching materials,
  -- lack of materials or equipment,
  -- lack of student interest in science,
  -- lack of teacher interest in science,
  -- teachers inadequately prepared to teach science,
  -- lack of support of administration,
  -- teachers' views not incorporated into curricular decisions, and
  -- lack of opportunity and/or support for inservice.

A Final Concern: Social Desirability

One legitimate concern that could be raised about the collection of questionnaire data for State-NAEP is the issue of social desirability. In Connecticut, we guaranteed complete anonymity--teachers and principals were told that their data would never be reported back to their schools. Therefore, we have every reason to believe that the responses we received were honest. If similar data are collected on State-NAEP and made public, we do not know whether different kinds of pressures will be brought to bear on students, teachers, and/or principals. If any of these groups is anything less than completely candid--or if there are systematic differences in the level of candor in different States--using data from these context variables will not only be useless; it will be misleading.

Summary

The ultimate criterion for inclusion on the National- or State-NAEP questionnaire should be whether an item will provide policymakers with information that will help them to improve education. No item should be included unless it has a strong history of providing useful policy-oriented data. Once included, an item should be used for several years so as to allow educators to monitor important changes which might be occurring within the educational context.

Questionnaires should be administered to students, teachers, and principals. For all three groups, "less is more." For State-NAEP to be accepted, the data collection burden must be perceived as minimal and reasonable and all questionnaire items should have "surface validity." The guidelines provided below have been developed with these criteria in mind. If a questionnaire item is related to student achievement:

- o does it place a student in one of the reporting categories predetermined as being important (e.g., gender, race, ethnicity, language spoken at home?); or
- o does it help to establish the test's concurrent validity? or
- o is it an alterable variable which meets these criteria:
  - -- it would be desirable for there to be more of this variable present, and
  - -- it is not easily corrupted and educators will attempt to monitor changes in closely related parts of the educational system.

If a questionnaire item is not related to achievement:

- o it should be a desired outcome of schooling, or it should be a desired belief or attitude resulting from schooling, or
- o it should be a necessary but not sufficient condition for improvement

A strong emphasis of the questionnaire should be on the alterable variables assessing opportunity to learn both the quantity of time spent in class and the quality of that time. Questions should relate to course-taking patterns, the amount of time spent on learning the subject both in school and outside of school (e.g., homework and long-term projects), the nature of the curriculum and its synchrony with the test, students' exposure to the tools and apparatus of the subject area, and the extent to which students report that they have internalized, and practice the behaviors of "experts."

Because of the pressures that may be felt with the advent of the State-NAEP, it would be advisable to pilot test the questionnaire items under conditions similar to those in which they will ultimately be administered.

References

Burstein, L (1980). Analysis of multi-level data in educational research and evaluation. In D. Berliner (Ed.), Review of Research in Education.

Brookover, W., Beady, C., Flood, P., Schweitzer, J., Wisenbaker, J., (1979). School social systems and student achievement: Schools can make a difference. New York: Praeger.

Connecticut State Department of Education. Connecticut Assessment of Educational Progress in Social Studies 1982-83 Technical Report. Hartford, CT.

Connecticut State Department of Education. Connecticut Assessment of Educational Progress in English Language Arts 1983-84 Technical Report. Hartford, CT.

Connecticut State Department of Education. Connecticut Assessment of Educational Progress in Science 1984-85 Technical Report. Hartford, CT.

Connecticut State Department of Education. Connecticut Assessment of Educational Progress in Foreign Language 1986-87 Technical Report. Hartford, CT.

Connecticut State Department of Education. Connecticut Assessment of Educational Progress in Drafting, Graphic Arts, and Small Engines 1986-87 Technical Report. Hartford, CT.

Forgione, P. D., Jr. and Baron, J. B. (1987). Can reporting on educational indicators serve as a catalyst for the improvement of educational achievement?--A visionary exploration. Presented at the annual meeting of the American Statistical Association, San Francisco.

Glaser, R. (1987). A speech given at conference on Subject Matter Assessment at Center for Research on Evaluation, Standards, and Student Testing, UCLA, December.

Jones, B. F., Palincsar, A. S., Ogle, D. S., Carr, E. G. (1987). Strategic teaching and learning: Cognitive instruction in the content areas. Alexandria, VA. Association for Supervision and Curriculum Development and the North Central Regional Educational Laboratory.

Koretz, D. (1988). Comments made during symposium linking policies with program features: Policy and technical issues in indicator construction for school reform assessment at Annual Conference of the Education Commission of the States and the Colorado Department of Education, Boulder, Colorado.

McKnight, C. G., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swatford, J. O., Travers, K.J. (1987). The underachieving curriculum: Assessing U.S. school mathematics from an international perspective. Champaign, IL: Stipes Publishing Company.

Murnane, R. J. (1987). Improving education indicators and economic indicators: The same problems? Presented at the annual meeting of the American Educational Research Association, Washington, D.C., April.

Oakes, J. (1987). Conceptual and measurement issues in the construction of school quality indicators. Presented in the annual meeting of the American Educational Research Association, Washington, D.C., April.

Ontario Ministry of Education (1988). Canadian Studies Geography - A Report for Educators.

Exhibits and Appendixes
(Baron and Forgione paper)

On all Exhibits the number to the right of each option is the average percent correct or mean test score (MTS) on the achievement test for the students choosing that option. The number in parentheses is the percentage of students selecting that option (PSO).

Exhibit 1

Measuring Students' Attitudes: Examples from Science (1984-85)

1. Careers in science are more appropriate for men than for women.

	Science 8		Science 11	
	(MTS)*	(PSO)**	(MTS)*	(PSO)**
strongly agree	41	(04)	42	(04)
agree	45	(10)	39	(07)
disagree	50	(36)	51	(44)
strongly disagree	53	(49)	53	(45)

2. My knowledge of science will be of little value to me in my day-to-day life.

	Science 8		Science 11	
	(MTS)*	(PSO)**	(MTS)*	(PSO)**
strongly agree	45	(09)	45	(08)
agree	48	(33)	46	(33)
disagree	53	(45)	53	(48)
strongly disagree	54	(13)	56	(11)

* (MTS) Mean Test Score
**(PSO) Percentage Selecting Option

Exhibit 2

Instructional Activities in
Connecticut Assessment of Educational Progress
In Social Studies (1982-83)

	Grade 8		Grade 11	
	(MTS)*	(PSO)**	(MTS)*	(PSO)**
1) How often does teacher lecture/students listen.				
never	53	(04)	47	(03)
a few times a year	58	(05)	61	(02)
at least once a month	60	(10)	60	(0-)
at least once a week	60	(36)	59	(21)
just about daily	56	(44)	60	(69)
2) How often do teacher/students discuss topics.				
never	40	(02)	61	(05)
a few times a year	49	(02)	63	(04)
at least once a month	56	(07)	59	(06)
at least once a week	58	(30)	60	(30)
just about daily	58	(59)	59	(55)
3) Teacher/students discuss current events in class.				
never	53	(07)	58	(08)
a few times a year	60	(07)	64	(09)
at least once a month	60	(21)	62	(25)
at least once a week	5-	(4-)	60	(40)
just about daily	54	(21)	54	(18)
4) Students express/defend opinions in class.				
never	53	(11)	59	(10)
a few times a year	56	(09)	65	(10)
at least once a month	60	(20)	58	(18)
at least once a week	58	(32)	60	(30)
just about daily	56	(29)	59	(32)
5) Teacher/students discuss TV programs in class.				
never	56	(34)	58	(22)
a few times a year	63	(26)	64	(28)
at least once a month	58	(25)	61	(29)
at least once a week	53	(13)	56	(17)
just about daily	49	(02)	49	(04)
6) Students do individual/group/class projects.				
never	49	(13)	55	(21)
a few times a year	61	(35)	63	(38)
at least once a month	58	(37)	61	(28)
at least once a week	56	(11)	53	(09)
just about daily	51	(05)	53	(03)

	Grade 8		Grade 11	
7) Students write reports.				
never	51	(13)	54	(18)
a few times a year	61	(43)	62	(49)
at least once a month	56	(37)	60	(27)
at least once a week	49	(06)	58	(05)
just about daily	40	(01)	41	(01)
8) Students choose topics of interest.				
never	56	(35)	59	(42)
a few times a year	61	(33)	65	(34)
at least once a month	58	(21)	58	(17)
at least once a week	49	(08)	47	(05)
just about daily	44	(04)	50	(03)
9) Students use library/media center for assignments.				
never	51	(17)	55	(22)
a few times a year	61	(27)	63	(37)
at least once a month	60	(33)	62	(24)
at least once a week	54	(17)	57	(13)
just about daily	49	(05)	46	(04)
10) How often do students/teacher meet individually.				
never	56	(52)	59	(56)
a few times a year	61	(22)	64	(23)
at least once a month	58	(13)	58	(13)
at least once a week	54	(08)	54	(06)
just about daily	53	(04)	48	(02)
11) How often do you have outside speakers visit.				
never	58	(53)	60	(58)
a few times a year	60	(37)	61	(35)
at least once a month	49	(05)	52	(05)
at least once a week	47	(03)	41	(01)
just about daily	33	(01)	29	(01)
12) How often are field trips.				
never	58	(62)	60	(75)
a few times a year	60	(34)	62	(21)
at least once a month	47	(02)	51	(02)
at least once a week	42	(01)	38	(01)
just about daily	42	(01)	43	(01)
13) How often do students use textbook in class.				
never	51	(04)	64	(17)
a few times a year	58	(04)	66	(12)
at least once a month	60	(14)	61	(17)
at least once a week	60	(40)	59	(30)
just about daily	54	(38)	54	(25)

	Grade 9		Grade 11	
14) How often are files/cassettes/videotapes used.				
never	53	(10)	57	(07)
a few times a year	58	(12)	62	(14)
at least once a month	60	(37)	63	(42)
at least once a week	58	(34)	57	(30)
just about daily	51	(06)	50	(06)
15) How often are games or models used.				
never	56	(56)	59	(68)
a few times a year	61	(24)	63	(20)
at least once a month	58	(13)	58	(08)
at least once a week	54	(05)	56	(03)
just about daily	54	(01)	44	(01)
16) How often are maps, charts, globes used.				
never	51	(05)	58	(08)
a few times a year	58	(11)	60	(16)
at least once a month	60	(27)	62	(28)
at least once a week	60	(33)	61	(29)
just about daily	54	(24)	55	(18)
17) How often are primary source materials used.				
never	56	(36)	60	(49)
a few times a year	60	(28)	64	(23)
at least once a month	58	(22)	60	(16)
at least once a week	54	(09)	52	(08)
just about daily	49	(05)	49	(04)
18) How often are computers used in class.				
never	58	(93)	61	(94)
a few times a year	56	(04)	47	(02)
at least once a month	47	(01)	37	(01)
at least once a week	44	(01)	45	(01)
just about daily	39	(01)	48	(01)
19) How often are multiple-choice tests given.				
never	58	(16)	57	(17)
a few times a year	60	(20)	62	(16)
at least once a month	61	(37)	63	(35)
at least once a week	51	(23)	57	(27)
just about daily	44	(04)	47	(05)
20) How often are students required to write own answers.				
never	49	(04)	59	(07)
a few times a year	56	(07)	61	(09)
at least once a month	61	(21)	64	(39)
at least once a week	56	(39)	57	(37)
just about daily	49	(09)	51	(09)

* (MTS) Mean Test Score
**(PSO) Percentage Selecting Option

Exhibit 3

Connecticut Assessment of Educational Progress
in English Language Arts (1983-84)

How often do you do each of the following to make your papers better?

	Grade 8		Grade 11	
	(MTS)*	(PSO)**	(MTS)*	(PSO)**
1) Move some sentences or paragraphs to different parts of the paper.				
almost always	5.4	(17)	5.1	(19)
more than half the time	5.2	(26)	4.7	(26)
about half the time	4.6	(24)	4.6	(22)
less than half the time	4.8	(24)	4.7	(18)
never or hardly ever	4.0	(09)	4.4	(12)
2) Add new ideas or information.				
almost always	5.2	(21)	5.1	(31)
more than half the time	5.0	(39)	4.8	(33)
about half the time	4.7	(24)	4.6	(18)
less than half the time	4.2	(10)	4.0	(15)
never or hardly ever	4.7	(06)	5.0	(01)
3) Take out parts of the paper that you don't like.				
almost always	5.5	(26)	4.9	(33)
more than half the time	5.0	(28)	4.7	(28)
about half the time	4.6	(27)	4.4	(24)
less than half the time	4.5	(10)	4.7	(05)
never or hardly ever	4.1	(08)	5.1	(07)
4) Change some words for other words that you like better.				
almost always	5.5	(27)	4.9	(38)
more than half the time	4.8	(27)	4.7	(33)
about half the time	4.7	(26)	4.7	(14)
less than half the time	4.5	(16)	4.1	(09)
never or hardly ever	4.3	(06)	4.6	(03)
5) Correct mistakes in spelling, grammar, and punctuation.				
almost always	5.0	(43)	4.7	(47)
more than half the time	5.1	(27)	5.0	(25)
about half the time	4.6	(19)	4.7	(15)
less than half the time	4.9	(10)	4.2	(09)
never or hardly ever	3.6	(02)	4.1	(02)
6) Rewrite almost all of the paper.				
almost always	5.1	(21)	4.2	(15)
more than half the time	5.0	(17)	4.4	(19)
about half the time	4.9	(16)	4.9	(19)
less than half the time	4.7	(25)	5.1	(26)
never or hardly ever	4.9	(22)	4.7	(20)
7) Throw out the first paper and start again.				
almost always	5.1	(08)	4.6	(09)
more than half the time	4.9	(12)	4.1	(09)
about half the time	4.6	(15)	4.7	(12)
less than half the time	4.6	(27)	5.0	(26)
never or hardly ever	5.2	(37)	4.7	(43)

* (MTS) Mean Test Score **(PSO) Percentage Selecting Option

Exhibit 4

Connecticut Assessment of Educational Progress
in English Language Arts (1983-84)

How often does your teacher ask you to do each of the following?

	Grade 8		Grade 11	
	(MTS)*	(PSO)**	(MTS)*	(PSO)**
1) Make notes before you write.				
almost every time	5.3	(21)	4.5	(22)
more than half the time	5.4	(16)	5.2	(19)
about half the time	4.5	(21)	4.6	(28)
less than half the time	4.9	(26)	4.9	(15)
never or hardly ever	4.1	(18)	4.6	(13)
2) Make an outline for the paper.				
almost every time	4.4	(09)	5.0	(22)
more than half the time	5.1	(12)	4.8	(12)
about half the time	5.1	(17)	4.7	(13)
less than half the time	4.8	(26)	4.9	(22)
never or hardly ever	4.8	(36)	4.4	(29)
3) Make notes for yourself about changes in the paper.				
almost every time	4.8	(23)	4.8	(20)
more than half the time	5.2	(16)	4.5	(15)
about half the time	5.1	(21)	4.7	(18)
less than half the time	5.2	(18)	4.9	(21)
never or hardly ever	4.0	(21)	4.7	(23)
4) Talk with the teacher about the paper while you are working on it.				
almost every time	4.9	(20)	4.4	(23)
more than half the time	4.9	(21)	4.9	(23)
about half the time	4.9	(27)	5.0	(22)
less than half the time	5.1	(17)	4.6	(19)
never or hardly ever	4.5	(16)	4.6	(10)
5) Talk with some classmates about the paper while you are working on it				
almost every time	4.5	(12)	4.4	(16)
more than half the time	4.7	(09)	5.3	(16)
about half the time	4.9	(18)	4.5	(19)
less than half the time	4.9	(24)	4.9	(22)
never or hardly ever	4.9	(37)	4.7	(25)
6) Write the paper more than once before it is graded.				
almost every time	5.1	(22)	4.5	(29)
more than half the time	4.9	(16)	4.8	(18)
about half the time	5.2	(14)	4.7	(16)
less than half the time	4.8	(22)	5.1	(21)
never or hardly ever	4.4	(26)	4.5	(13)

	Grade 8		Grade 11	
	(MTS)*	(PSO)**	(MTS)*	(PSO)**
7) Work on the paper again after it has been graded.				
almost every time	4.5	(09)	4.6	(06)
more than half the time	5.2	(09)	4.5	(05)
about half the time	4.1	(14)	4.7	(15)
less than half the time	5.3	(22)	4.9	(22)
never or hardly ever	4.8	(47)	4.7	(50)
8) Use a dictionary or thesaurus.				
almost every time	5.1	(25)	4.6	(31)
more than half the time	5.0	(21)	4.9	(25)
about half the time	4.8	(22)	4.9	(15)
less than half the time	4.6	(21)	4.7	(14)
never or hardly ever	4.4	(11)	4.6	(13)
9) Reread your writing to yourself.				
almost every time	5.0	(51)	4.9	(53)
more than half the time	4.8	(19)	4.5	(24)
about half the time	4.6	(12)	4.4	(13)
less than half the time	4.8	(11)	5.0	(04)
never or hardly ever	4.1	(05)	4.7	(04)
10) Read your writing to someone else.				
almost every time	5.1	(14)	4.8	(16)
more than half the time	5.4	(14)	5.0	(17)
about half the time	5.2	(15)	4.9	(13)
less than half the time	4.6	(19)	4.7	(25)
never or hardly ever	4.5	(37)	4.4	(27)

* (MTS) Mean Test Score
**(PSO) Percentage Selecting Option

Exhibit 5

Course Taking Patterns: An Example from Science (1984-85)

o How many years of high school science will you take?

	Grade 8 (MTS)*	(PSO)**
One	42	(9)
Two	47	(30)
Three	53	(27)
Four	55	(34)

* (MTS) Mean Test Score
**(PSO) Percentage Selecting Option

Exhibit 6

Course Taking Patterns: Examples from Science 1984-85 (Grade 11)

o In grades 9-12, have you taken or do you intend to take each of the following science courses? (The two sets of data correspond to two forms of the questionnaire which were matrix sampled. We have provided both sets of responses which can be viewed as "cross-validation samples.")

	Have Taken or am Currently Taking		Intend to Take Next Year		Do Not Plan to Take	
	(MTS)*	(PSO)**	(MTS)*	(PSO)**	(MTS)*	(PSO)**
General Science	50	(62)	35	(02)	52	(36)
	50	(63)	31	(02)	54	(36)
Earth Science	53	(57)	45	(03)	48	(40)
	5[illegible]	(57)	40	(02)	50	(40)
Biology	52	(87)	37	(04)	41	(10)
	53	(86)	37	(03)	35	(10)
Chemistry	58	(50)	46	(13)	43	(37)
	59	(49)	49	(13)	41	(37)
Physics	51	(16)	60	(25)	47	(59)
	55	(18)	60	(24)	45	(58)
Second-year Biology***	52	(06)	54	(13)	50	(82)
	51	(06)	57	(12)	50	(82)
Second-year Chemistry***	49	(03)	51	(07)	51	(91)
	50	(02)	55	(05)	5[illegible]	(93)
Second-year Physics***	39	(02)	50	(08)	51	(90)
	45	(02)	51	(06)	51	(92)

* (MTS) Mean Test Score
**(PSO) Percentage Selecting Option

***There may have been a misunderstanding about the meaning of "Second-Year" --it looks like the "repeaters" have included themselves in the "Have Taken or am Currently Taking Group." We probably should have used the word "Advanced."

Exhibit 7

Course Taking Patterns of Males and Females in (1984-85)
Science Coursework of Grade 11 Students

Percent of Students Having
Taken, Taking, or Intending
to Take Course

Course	Males	Feamles	All
General Science	69	61	65
Earth Science	64	53	59
Biology	83	94	89
Chemistry	66	58	62
Physics	54	30	42

Exhibit 8

Students' Performance, and Teachers' and Students' Ratings of Students' Competence on the Eight Domains of the Drafting Multiple-Choice Test

Teachers' Ratings of Students' Competence			Students' Ratings of Own Competence			Test Domain	Number of Test Items in Domain		Percentage Correct	
Clearly Above or Just competent	Not Competent	Domain Not Covered	Clearly Above or Just competent	Not Competent	Domain Not Covered		Total Test	Critical Items	Total Test	Critical Items
90%	8%	2%	78%	13%	10%	Machine Drawing	24	9	43%	45%
78%	17%	5%	84%	7%	9%	Architectural Drawing	18	10	56%	69%
93%	5%	2%	86%	11%	3%	Interpretation of Drawing	46	32	55%	64%
17%	22%	61%	33%	25%	42%	Electronic Drawing	5	3	46%	61%
82%	16%	2%	80%	11%	8%	Related Mathematics	12	11	60%	61%
48%	30%	24%	39%	19%	42%	Sheet Metal Drawing	5	0	46%	--
15%	17%	68%	33%	19%	48%	Mapping	5	0	48%	--
44%	29%	27%	52%	17%	31%	Computer Drawing	5	1	39%	55%
						TOTALS	120	66	52%	61%

Exhibit 9

Amount of Experience with Apparati and Technology:
Example from Science 1984-85 (Grade 4)

		(MTS)*	(PSO)**
1.	Have you ever used a magnifying glass in science?		
	A. Yes	55	(44)
	B. No	52	(56)
2.	Have you ever used a metric ruler in science?		
	A. Yes	53	(48)
	B. No	53	(52)
3.	Have you ever used a thermometer in science?		
	A. Yes	55	(38)
	B. No	52	(62)
4.	Have you ever used a magnet in science?		
	A. Yes	53	(39)
	B. No	52	(61)
5.	Have you ever made a simple electrical circuit?		
	A. Yes	53	(32)
	B. No	52	(68)
6.	Have you ever made an electromagnet in science?		
	A. Yes	53	(12)
	B. No	53	(88)

* (MTS) Mean Test Score
**(PSO) Percentage Selecting Option

Exhibit 10

Students' Experience with Laboratory Equipment
Science (1984-85)

Experience with: Percent at:	Triple-Beam Balance Gr. 8	Gr. 11	Graduated Cylinder Gr. 8	Gr. 11	Setting up Electrical Circuit Gr. 8	Gr. 11	Microscope Gr. 8	Gr. 11
Never	29	17	11	4	29	29	3	6
1 or 2 times	23	12	25	9	25	21	13	3
3 to 5 times	17	10	22	12	19	12	18	5
6 to 10 times	12	10	13	12	9	13	24	13
More than 10 times	18	49	28	62	16	22	41	69

Previous Experience and Performance
With Triple-Beam Balance

Grade 8 Experience with Triple-Beam Balance

	0-2 times	3 or more times
Measured acceptable	29	48
Weight of block not acceptable	49	23

Exhibit 11

Connecticut Assessment of Educational Progress
in English Language Arts (1983-84)

	Grade 4		Grade 8		Grade 11	
	(MTS)*	(PSO)**	(MTS)*	(PSO)**	(MTS)*	(PSO)**
1) About how often do you presently use a computer in school?						
every day or almost every day	30	(05)	43	(04)	64	(11)
a few times a week	35	(16)	46	(13)	67	(10)
a few times a month	31	(18)	51	(09)	67	(06)
once a month or less	30	(13)	49	(11)	55	(14)
never	24	(46)	36	(63)	44	(56)

	Grade 4		Grade 8		Grade 11	
2) How many times ever have you used a school computer to learn or practice computer programming?						
more than 20 times	36	(05)	48	(11)	72	(21)
11 to 20 times	37	(05)	58	(06)	61	(04)
6 to 10 times	34	(07)	42	(09)	52	(05)
1 to 5 times	29	(20)	48	(17)	57	(10)
never	26	(60)	33	(56)	43	(58)

* (MTS) Mean Test Score
**(PSO) Percentage Selecting Option

Exhibit 12

Connecticut Assessment of Educational Progress
in English Language Arts"(1983-84)

	Grade 8 (MTS)*	(PSO)**	Grade 11 (MTS)*	(PSO)**
1) How much time do you spend doing all of your homework assignments on a typical school day?				
less than half an hour	51	(14)	69	(21)
half an hour to 1 hour	59	(40)	74	(33)
between 1 and 2 hours	64	(36)	79	(35)
between 2 and 3 hours	65	(09)	80	(09)
more than 3 hours	47	(02)	72	(01)
2) How much time spend studying for important for important test?				
less than half an hour	50	(14)	67	(13)
between half an hour and 1 hour	61	(41)	73	(33)
between 1 and 2 hours	64	(34)	78	(34)
between 2 and 4 hours	57	(08)	80	(15)
more than 4 hours	47	(02)	92	(01)
3) How often do you read material over a few times?				
almost every time	63	(38)	81	(33)
more than half the time	63	(26)	77	(25)
about half the time	54	(23)	69	(24)
less than half the time	53	(10)	66	(12)
never or hardly ever	49	(02)	78	(04)
4) How often do you read parts of a novel or story?				
almost every day	69	(13)	80	(17)
once or twice a week	60	(28)	78	(29)
once or twice a month	64	(25)	71	(27)
a few times a year	57	(16)	67	(16)
never or hardly ever	45	(16)	54	(09)

* (MTS) Mean Test Score
**(PSO) Percentage Selecting Option

Exhibit 13

Asking Students About "Press for Achievement":
Examples from Science (1984-85)

1. How many students in this school try hard to get good grades on their classroom tests?

	Grade 8 (MTS)*	(PSO)*
Almost all of the students	49	(18)
Most of the students	53	(47)
About half of the students	52	(25)
Some of the students	46	(10)
Almost none of the students	50	(01)

2. How many students in this school don't do as well as they could in school because they are afraid their friends won't like them as much?

	Grade 8	
Almost all of the students	41	(02)
Most of the students	42	(06)
About half of the students	47	(11)
Some of the students	53	(42)
Almost none of the students	52	(40)

3. How often do you come to science class with all the materials you need? (e.g., pens, paper, books, etc.)

	Grade 8	
Always	52	(63)
Most of the time	52	(31)
About half of the time	46	(04)
Once in a while	44	(01)
Never		(00)

4. How often do you put a lot of effort into your science homework?

	Grade 8	
Always	51	(26)
Most of the time	53	(53)
About half of the time	47	(14)
Once in a while	43	(06)
Never	45	(01)

5. My teacher cares about how well I do in Science

	Grade 4
Usually true	54 (83)
Sometimes true	50 (14)
Almost never true	43 (03)

	Science 8	Science 11
Strongly agree	49 (32)	52 (18)
Agree	53 (55)	52 (57)
Disagree	49 (10)	50 (19)
Strongly disagree	47 (04)	41 (06)

* (MTS) Mean Test Score
**(PSO) Percentage Selecting Option

Appendix A

Educational Resources

Percentage of Teachers Selecting Each Option - Connecticut Assessment of Educational Progress in Science (1984-85)

1. In general, how many students in this school try hard to get good grades?

	Science 4	Science 8	Science 11
almost all of the students	14	07	03
most of the students	52	48	26
half the students	23	32	39
some of the students	10	13	28
almost none of the students	0	1	2

Percentage of Principals Selecting Each Option - Connecticut Assessment of Educational Progress in Science (1984-85).

2. How much difficulty do you have in securing qualified science teachers to fill vacancies?

	Grade 11
a great deal of difficulty	42
some difficulty	42
little or no difficulty	14

3. Does your school have a petty cash fund that can be used for science supplies?

	Grade 8	Grade 11
Yes	52	52
No	48	48

215

4. How much does your school annually budget specifically for the purchase of consumable science supplies (materials that must continually be replenished such as chemicals, glassware, batteries, etc.)?

	Science 4	Science 8	Science 11
$100 or less	17		
$101-$200	9		
less than $200		6	
$201-$400	29	9	
$401-$600	14	9	
less than $600			3
more than $600	29		
$601-$800		6	
$801-$1,000		8	
$1,001-$1,200		5	
$600-$1,200			12
$1,200-$1,400		9	
$1,401-$1,600		9	
more than $1,600		35	
$1,201-$1,800			10
$1,801-$2,400			13
more than $2,400			59

5. How much does your school annually budget specifically for the purchase of new science equipment (nonconsumable, nonperishable items such as microscopes, scales, etc.--not textbooks)?

	Science 4	Science 8	Science 11
$100 or less	17		
$101-$200	24		
$200 of less		11	
$201-$400	24	15	
$401-$600	11	11	
less than $600			10
more than $600	23		
$601-$800		13	
$801-$1,000		9	
$1,001-$1,200		13	
$600-$1,200			13
$1,201-$1,400		8	
$1,401-$1,600		4	
more than $1,600		14	
$1,600-$1,800			13
$1,801-$2,400			19
more than $2,400			41

6. How many microcomputers does your school have for students use related to science instruction?

	Science 4	Science 8	Science 11
0	35	12	19
1 or 2	21	16	26
3-5	15	10	17
6-10	15	11	7
more than 10	13	29	30

7. Are your students homogeneously grouped?

	Science 4	Science 8	Science 11
Yes	15	57	80
No	83	43	20

Appendix B

Connecticut Assessment of Educational Progress
in Science 1984-85

The following factors may affect science instruction, and ultimately achievement in your school as a whole. In your opinion, how much of a problem is caused by each of the following?

Percentage of Teachers (T) and Principals (P)
Choosing Each Level of Problem

	Grade Level	Serious Problem		Somewhat of a Problem		Not a Significant Problem	
		T	P	T	P	T	P
a general belief that science is less important than other subjects	04	15	06	45	53	39	37
	08	09	00	32	11	57	89
	11	12	01	31	19	56	80
out-of-date teaching materials	04	20	05	32	21	47	69
	08	11	00	33	19	55	81
	11	13	04	34	36	51	59
lack of materials or equipment	04	28	12	36	37	35	47
	08	17	03	32	27	51	71
	11	20	13	39	35	40	52
inadequate budget for science	04	23	11	42	28	32	57
	08	22	06	36	22	42	72
	11	32	10	39	36	28	54
lack of student interest in science	04	07	00	31	19	59	76
	08	13	00	39	25	48	75
	11	21	04	49	33	29	62
lack of teacher interest in science	04	17	10	34	49	45	36
	08	06	03	15	15	78	82
	11	05	00	16	09	77	91
teachers inadequately prepared to teach science	04	20	21	40	50	36	25
	08	10	04	25	22	64	75
	11	07	01	22	14	69	84

	Grade Level	Serious Problem		Somewhat of a Problem		Not a Significant Problem	
		T	P	T	P	T	P
lack of support of administration	04	14	02	31	11	51	80
	08	14	01	32	14	54	84
	11	23	01	38	22	39	75
teachers' views not incorporated into curricular decisions	04	14	01	41	17	41	75
	08	12	00	33	10	54	90
	11	17	04	31	10	50	84
lack of opportunity and/or support for inservice	04	19	05	35	31	39	56
	08	11	03	38	49	51	48
	11	18	10	38	38	42	51

Appendix C

Connecticut Assessment of Educational Progress
in Science 1984-85

Teachers

1) How available is science equipment (e.g., hands-on materials, glassware, chemicals) for your use in teaching science?

	Science 4	Science 8	Science 11
I have all I need without having to share with other teachers.	18	35	30
I must share with other teachers in the school to get what I need	53	56	53
I must borrow equipment from another school (e.g., high school) to get what I need	5	0	NA
I cannot acquire a a lot of the equipment that I need.	21	8	16

2) How well trained are you to teach science at the level you teach?

	Science 4	Science 11
not well trained at all	12	02
adequately trained	64	22
very well trained	22	76

Appendix D

The Percentage of Teachers
Who Report That the Target Foreign Language
Is Used in Courses 2-6
of the Modern Languages

	Mostly English	Equal Amounts of English and this Foreign Language	Mostly this Foreign Language	Onlythis Foreign Language
Teacher speaks				
Course 2	6	53	7	3
Course 3-6	1	16	64	19
Students speak				
Course 2	25	52	21	2
Course 3-6	4	30	55	11
Students hear				
Course 2	5	52	40	3
Course 3-6	1	15	67	17
Students write				
Course 2	1	15	46	38
Course 3-6	0	4	31	64

Reporting State-Level NAEP in a Fair and Credible Manner

Leigh Burstein
University of California, Los Angeles

According to current plans as articulated in legislation, the 1990 NAEP will collect State-representative data in up to 30 States wishing to participate at grade eight in mathematics; the 1992 NAEP will expand the trial to reading and mathematics at grade 4. Beginning in 1994, data collection would be expanded to all subject areas and grade levels. To prepare for the 1990 effort, the National Center for Education Statistics (NCES) contracted with the Council of Chief State School Officers (CCSSO) State Education Assessment Center to model the Consensus Planning process, identified in the Alexander James report (Alexander and James, 1987) as the new way to govern NAEP, and in the process, to develop plans for the 1990 trial in mathematics (content specifications, analysis and reporting guidelines). In addition, the NAEP Technical Review Panel was asked to make recommendations on the same set of issues.

One major task in planning and conducting State-level NAEP is to identify critical issues and analyze options in what to report and in what form(s) to report it. The intent of this paper is to provide a framework for and present the panel's recommendations regarding the reporting of State-level NAEP data collected during the 1990 and 1992 pilot efforts. Experiences with State-level NAEP reporting during the pilot efforts should serve as a starting point for decisions about such reporting in anticipated fully operational State-level NAEP data collection beginning with the 1994 National-NAEP.

Other papers prepared by this panel review administrative procedures and matters regarding the contents of the assessment as implemented in State-level data collection. The recommendations from companion papers establish the standards for the conduct of State-level NAEP in areas other than the reporting of State-level data. In addition, the report "Within-State Comparisons: Suitability of State Models for National Comparisons" prepared by Haertel as an activity separate from the panel's work but included as part of its final report, examines specific methods currently used by various States for presenting comparisons among schools or districts within a State and their applicability for between-State comparisons. Since Haertel's report already covers much of the ground regarding the rationale, mechanics, and strengths and weakness of options currently employed around the country, the focus here is on the broader issues that motivate and could guide the reporting of State-level NAEP data.

Recommendations on State-Level Reporting

In the final report submitted by the NAEP Technical Review Panel, two recommendations, and their accompanying rationales, explicitly dealt with the reporting of State-level data:

RECOMMENDATION 10

The expansion of NAEP to provide data at the level of individual States will entail careful study of methods for making and reporting State comparisons. In the 1990 and 1992 pilot studies, a variety of methods should be explored and reported.

Where feasible, State results should be reported for major process and content categories, using the same proficiency scales as are used for National-NAEP. In many content areas, age-specific proficiency scales may be more useful and appropriate than scales spanning different age/grade levels. In addition to reporting absolute levels of achievement on these scales, each State's performance might be referenced to that of a small group of comparable States, or to nationally representative samples of students matched to State population characteristics. Additional alternatives may also be explored.

RECOMMENDATION 11

The reporting of cross-sectional and trend results for State-level NAEP should characterize both the level and distributions of student attainment within each State. This reporting should include (a) demographic subgroup and community differences; (b) variation in performance across major domains of learning outcomes; and (c) distributions of school-level performance within the State.

Reporting score distributions for major subdomains is more informative than reporting means for broad content areas. This is true at the State level as well as the national level. State and national score distributions for major subdomains should be reported in ways that facilitate their direct comparison to one another.

In addition to distributions for entire States, performance should be reported for demographic subgroups and types of communities within States, whenever such reporting is feasible. Feasibility may be limited by smaller sample sizes for groups or areas within States, or by legal requirement that results not be reported for schools or districts in the 1990 and 1992 pilot assessments.

Because the school is an important locus of educational policy, we recommend that distributions of school means as well as distributions of individual scores be reported. Where samples of schools are sufficiently large and representative, distributions of school means should be reported for States, and for different types of schools within States. By law, particular schools would not be identified.

These recommendations have much in common with those produced by the CCSSO NAEP Assessment Planning Project (CCSSO, 1988) on matters of reporting. In part this can be attributed to the overlapping memberships of the two groups (both Forgione and Burstein served on the CCSSO committee that dealt with analysis and reporting matters). But perhaps more important was a shared belief in both projects that the 1990 and 1992 trials represent opportunities to offer alternatives to existing practice in depicting educational performance across the States that could set a new standard for quality and comprehensiveness. Given multiple audiences for State-level NAEP and their potentially competing political and educational agendas, there are obvious risks in attempting to portray the accomplishments of the students from individual States. The hope is that by providing reports that fully and faithfully characterize performance at the State level in a variety of diverse ways, the likelihood of simplistic and misleading inferences will be reduced and the possibility of informed dialogue about the status of education across the Nation will be enhanced. Moreover, the resulting debates about the meaning of the various reporting alternatives provide the proper atmosphere in which to hone plans for reporting if and when State-level NAEP were to be fully implemented.

In the jargon of "evaluation utilization," while the explicit goals and purposes of these two efforts are the same, to a certain extent, their "clients" differed. CCSSO, either explicitly or implicitly, attempted to reflect the consensus views of the States as they examined this presumably national question. The NAEP panel, on the other hand, attempted to represent a perhaps broader set of constituents, the most critical of which is some notion of the "Nation's best interests" or the national public good. In a fundamental way, the groups were serving the same constituents, but there are points where the burden of choice dictated different decisions given the inherent resource constraints and somewhat different primary audiences. In what follows, then, this tension is acknowledged but not emphasized. Instead, the broader frameworks that served as the basis for both sets of recommendations on State-level NAEP reporting are discussed and key issues and options are highlighted.

Background

A starting point for most discussions about possible methods for analyzing and reporting State-level NAEP is to consider methods currently employed by many States and large city school districts. As Haertel recounts, a variety of procedures have been used by States to report and compare student achievement among schools and districts. The technical ancestry of the statistical and psychometric methods employed can be traced either to research on educational productivity and school effects conducted primarily by economists and sociologists (e.g., Averch et. al., 1972; Coleman et al., 1966) in the 1960s and early 70s, or to the literature on identifying unusually effective schools and school effectiveness (e.g., Brookover et. al., 1979; Edmonds, 1979; Klitgaard & Hall, 1974) that developed to some degree in response to the negative results from the earlier studies. Despite ongoing debates about the technical adequacy of the various analytical methods (e.g., Dyer et al., 1969; Mosteller &

Moynihan, 1969; Purkey & Smith, 1983), by the mid-1970s, a number of States were either conducting school effectiveness studies of their own (e.g., California State Department of Education, 1977; New York State Department of Education, 1977) or reporting school and district performance results.

The State-based studies and reporting systems of the 1970s were responses to the first wave of politically mandated educational accountability following the rapid expansion of educational programs and services in the war on poverty. Such efforts mounted by States molded the assessment capacity that developed following the introduction of NAEP and the diffusion of its technology to the State level with program evaluation expertise derived from efforts to evaluate the effectiveness of compensatory education. While most States retained and refined their assessment systems, and to some degree, their policy analytic functions over the years (Burstein et al., 1985), the reform movement of the 1980s generated by the report of the Excellence Commission and the associated spotlight from the publication of the Wall Chart (e.g., U.S. Department of Education, 1984) led to a major overhaul and expansion of State-level accountability and assessment activities to both stimulate school reforms and monitor their progress. And, while there have been refinements in the terminology (e.g., quality assessment, quality indicators, report cards; e.g., California State Department of Education, 1986) and in the comprehensiveness (e.g., assessment at more grade levels and reporting information in addition to achievement) and attractiveness of the reports (better graphics, higher quality printing), the technical and analytical underpinnings of State achievement monitoring systems, and associated complications with their use, remain much the same as before. States are certainly wiser in realizing the multifaceted nature of schooling, recognizing the complexity of assessing the impact of reforms and thus the futility of analytical "quick-fixes." Yet, expectations of the policy and practice communities about the documentation of the consequences of reform are stronger than ever. Moreover, the level of public trust is such that protestations that "tests weren't intended to serve these purposes" and "it's not technically feasible to assess reform impact" are politically unacceptable.

Just as State agencies have felt compelled to respond to the changing climate for educational information brought about by both an expanded audience and changing conditions in the Nation's educational systems, national organizations of State-level political and educational officials have joined the effort to improve the information base for monitoring educational progress (CCSSO, 1984, 1987, 1988; National Governors Association, 1987). These organizations have expressed their dissatisfaction with existing federal data sources on student achievement (primarily the Wall Chart reporting of performance on college admissions tests at the State level) and, after consideration of the alternatives, have backed the expansion of NAEP to provide State-level data. These organizations are not lending their support without regard to concerns about the quality of the reporting. The system they envision (e.g., CCSSO, 1988; Selden, 1986) is both a comprehensive and credible one that would inform educational debates within participating States.

Purposes for Reporting State-Level NAEP

A fundamental reason for State-level NAEP is a belief that when used appropriately and carefully, high quality information disaggregated to levels of authority for educational governance can help improve education. This belief justifies the development of a system to collect State-level data, but does not mean that the use of such data should be limited to gross, simplistic State comparisons of the kind often seen with comparative school achievement data (e.g., U.S. Department of Education, 1984). That is, one must avoid league tables and "wall charts" that simply depict raw rankings on mean achievement. Instead, the availability of State-level NAEP should enable the development of "wall pictures" that capture the full array of challenges faced by the Nation's educational systems along with their accomplishments and shortcomings.

Comparisons of educational entities at any level of aggregation are inherently judgmental. To minimize the invidious aspects of reporting of comparisons of performance based on State-level NAEP, one must acknowledge that States differ in the economic and demographic settings in which schools operate and in the social and economic background of the students entering the schools. Yet prevailing differences among States along these dimensions cannot be construed as a reasonable excuse to perpetuate inequalities in educational expectations, opportunities, or results. Rather, that socio-demographic factors may account for part of the outcome differences among the States, and the students within them, should be construed as a means to accentuate the magnitude of the task to achieve equal results and to help pinpoint areas of progress or regress.

Once the above points are acknowledged, a variety of ostensible purposes justify the conduct of State-level NAEP. One purpose, articulated early on within the CCSSO NAEP Assessment Planning effort, is as follows:

> Describe and monitor the condition of educational achievement in the Nation with respect to the States to inform and focus deliberations at the State policy level regarding the improvement of educational performance.

The Steering Committee for the CCSSO Project also identified five conditions that should be met for State-comparative data to be useful:

- They represent performance on a consensus of what is important to learn;
- They are based on sound testing and psychometric practice;
- They are based on procedures that minimize intrusion into instructional time;
- They account for the different circumstances and needs that the States face; and
- They are tied to concrete features of the school systems that can be changed for the better by State and local educators.

Other purposes could be identified and discussed, but these CCSSO conditions are sufficiently broadly Stated to serve as a basis for judging

the validity, credibility, and utility of State-level NAEP for purposes beyond those envisioned by State-level constituencies.

At the same time, however, there also needs to be consideration of how tradeoffs in implementation due to resource constraints and other factors impact the relative ability to address the various purposes. For example, the NAEP Technical Review Panel concluded that to represent important learning domains adequately while at the same time gathering information about the educational processes and practices to relate to performance, would likely entail an increase in student testing time. But expansion of the span of content (and, correspondingly, the number of test items administered) and other data collection adds both fiscal costs and respondent burden unless other intrusions on instructional time are minimized (such as reducing or eliminating other assessment and data collection activities in response to district, State, or federal requirements).

Hopefully, an exchange of State-level NAEP data collection and reporting for some other data collection activity will be warranted by the quality and utility of the information generated through NAEP.

Guiding Principles in Analysis and Reporting

Given the purposes articulated above, there are a number of principles that should guide the development of the design of State-level NAEP reporting.

A. Fundamentally, the State-level data should reflect what a State's students know and don't know and for a given time-frame, what they have learned and haven't learned. Stated in another way, the State-level data system should be designed to monitor progress as well as status. As such a long-term goal of data collection should be the reporting of State-level trend data.

B. The distribution of performance within the State is as important or more important than a raw State central tendency. Either percent within the State in national -iles (e.g., quartiles, deciles) and proportions in categories defined by some standard setting process (e.g., variations of NAEP's proficiency scales (NAEP, 1985, 1988)) are desirable options. In the latter case, age or grade-specific proficiency scales may be more informative than scales that span all age/grade levels. Moreover, the number of classifications on the scale should be sufficient to clearly demarcate distinctive levels of functioning to monitor progress over time.

C. It is better to compare subgroups within a State across States (i.e., black females) than to present gross central tendencies for all States. The more refined the subgroups (i.e., white males whose parents have less than high school education), the more informative the reporting. This principle holds to the degree that equity in performance is a consideration, assignments to subgroups can be accurately made, and estimates of performance at the subgroup level can be made with sufficient precision.

D. Reporting performance comparisons based on complex technical adjustments (e.g., differences between raw, unadjusted scores and expected scores derived from regressions of performance on background characteristics; see Haertel's paper) is inadvisable if for no other reason than the units for such comparisons are the 50 States. If comparisons are to be made, scores should be reported in the original raw metric units.

E. Comparisons of "comparable" States should employ classification procedures that result either in explicit categorizations whose basis is easily detectable (wealth rankings, regions) or a continuum derived from such data.

F. Gross aggregations of content/process in a learning domain can cause bad signals to political and educational leaders resulting in particularly invidious assessment impact. The measurements used to generate scores should be sufficiently disaggregated to detect content and process trends. For example, in mathematics score scales should not confound performance trends in computational proficiency with problem solving abilities nor those in basic arithmetic operations with understanding and application of algebraic relationships.

G. Although the design of a State-level data system is the target, the data collection and reporting should be sensitive to heterogeneity in performance across districts, schools, and classrooms as well as students. Measures of the variability of performance among the educational units at a given system level represent a desirable adjunct to State-level reports. A corollary of this principle is that the number of districts, schools, and classes obtained in the samples from each State should be sufficient to allow for such reporting.

H. Whatever basis is used to define the applicable population of students in a given State should be consistently employed across States in State-level data reporting. This means that definitions of excluded populations (e.g., special education, limited English proficiency, private schools) should be common across States for reporting purposes. Differences in the proportions of students excluded across States in various categories should be reported to further contextualize the results.

I. The samples drawn from each State should be sufficiently large to represent all demographic groups that are of interest nationally but only for those subgroups of sufficient size in the given State. Thus, no attempt should be made to oversample in States where particular groups represent a minute fraction of the student population (e.g., Hispanics in North Dakota). Samples should be State representative demographically and a minimum sample proportion established below which a subgroup's data are not reported.

Comparisons of States to What?

While a number of reports and organizations call for State-level comparisons, they are less clear as to the kinds of comparisons of greatest

interest. A number of distinctive options are possible. It is likely that several will be needed to satisfy the conditions implied by the guiding principles while achieving the purposes identified earlier.

The final recommendations from the panel, like those from the CCSSO project, focus on two distinctive types of comparisons: reporting student achievement data for each State separately (recommendation 11) and reporting comparisons among States (recommendation 10). With respect to the former, we envision the production of a series of "State data pages" wherein the performance of each participating State's sample on NAEP would be reported according to the same principles that govern the reporting of national results. The panel's recommendations for National-NAEP call for more detailed and informative reporting through greater specificity in representing domains of learning (recommendation 8) and reporting distributions of performance overall and by demographic factors (recommendation 9). In addition, we call for reporting the distribution of school-level performance within each State as yet another way to portray the homogeneity or heterogeneity of results. Such distributions are a useful means to characterize the disparities in performance among schools within each State's purview.

With the exception of the guiding principles that directly address this issue, we are less secure in proposing specific choices among the methods for State-to-State comparisons, including those discussed by Haertel. What is clear, however, is that most of the alternatives that have been proposed arouse strong feelings. In a survey of State educational and political officials conducted as part of the CCSSO project, roughly half of the respondents strongly advocated straightforward rankings of States on raw performance scores and violently objected to any attempts to adjust scores in any way, and the other half expressed exactly the opposite set of preferences. Similarly, opinions regarding alternatives for taking into account differences in regional and demographic conditions were diverse and strongly held.

Given the circumstances described above, there are a number of options that should be considered for the 1990 and 1992 trials and would likely provoke considerable interest and attention. Below is the manner in which several options might be operationalized is discussed briefly.

State to Nation

In envisioning a reporting system that most directly serves the interest of all States who choose to participate, the availability of information that allows a given State to contrast its performance with that of National-NAEP seems least controversial. Here whatever metrics, scales, and subgroups are deemed desirable are used to produce essentially a unique State-level data "page" with comparisons to the Nation as a whole. This comparison treats National-NAEP results as a standard for judging each participating State's performance.

The seemingly straightforward and seductive simplicity of this method of reporting should not lead one to lose sight of the fact that the

circumstances in a given State might be very different from those in the Nation as a whole. National results are derived from a sample that is nationally representative with respect to demographic stratification factors such as race/ethnicity and community type. When national results are disaggregated to the specific cells in the stratification, differences in performance are routinely observed (e.g., NAEP, 1985, 1988; Southern Regional Education Board, 1987). Given such patterns, one should not assume, for example, that it is reasonable to project the aggregated State-level performance in Mississippi with its high poverty rates and substantial minority population onto the national results with distinctively different concentrations of these demographic conditions. Yet, to interpret performance differences typically associated with demographic factors as inevitable, and thus as a basis for lowering performance expectations for a given State, would lead to implicit acceptance of conditions that might arise as the result of discriminatory educational practices.

The dilemma posed above cannot be resolved by choosing between total sample and substrata specific comparisons. Either both or neither are meaningful. The combination of State substrata comparisons to corresponding substrata nationally with State totals to the Nation as a whole is more informative and less risky.

Nation to State

Given the inherent complexity of State-to-Nation comparisons, one could instead attempt to project what the national performance would look like if the Nation's student and school demography were like that of the State. Thus for each participating State, separate national estimates could be derived wherein the weights applied to derive estimates of "State-adjusted national performance" are determined by the characteristics of the State's student population.

Just as the State-to-Nation comparison has limits, the Nation-to-State does as well. Without deep stratification by demographic variables (multiply cross-classified), the comparability of the national cells to the supposedly corresponding State cells is suspect. Moreover, the disparity in sample sizes between national and State results can lead to widely divergent sampling errors in estimating performance levels at the substrata level. Subsequent aggregation of both national and State results across substrata to compare performance of Nation and State would mask the differential instability of cell estimates.

State to "Pseudo-State"

A possible compromise that reflects the same intent as the Nation to the State but can potentially avoid some of the potential invidiousness of State-to-State comparisons is to attempt to construct a meaningful comparison group for each State from either the National-NAEP sample, the pooled sample encompassing data from all States participating in State-level NAEP, or both. One means would be to use the national data pool to construct a distribution of "pseudo-States" for each State based on that

State's demographical distribution. Operationally, the comparison to "pseudo-States" might proceed as follows:

1. Assign both national and State data to demographic substrata.

2. Select cases from each substratum in the national sample with frequency equal to number of cases in the State within that substratum. The resultant sample has the size and demographic makeup of the State's NAEP sample and thus represents a demographically comparable "pseudo-State."

3. Estimate performance statistics using the pseudo-State sample data.

4. Repeat Steps 2 and 3 a sufficient number of times (e.g., 100 replicates) to generate the sampling distribution of the performance statistics from the State-size, pseudo-State samples. The sampling variance of the distribution would depend on the properties of the State's demography.

5. Locate the State's actual performance within the sampling distribution constructed from the pseudo-State estimates. For example, if 100 such pseudo-State samples were drawn, one could place the State's own performance at the ?%ile of this distribution.

The State-to-pseudo-State comparison can only be as good as the demographical breakdown used to classify students. When the State's data are assigned to cells based on multiply cross-classified demographic factors, the correspondence of the sample from the national pool with that from the State is enhanced if cases are reliably classified. But the reliability of classification is highly dependent on the quality of the measurement of student demographic characteristics and on the number of stratification factors employed in the cross-classification. Moreover, as the number of stratification factors increase, the cell frequencies in the cross-classification of State data become less stable, introducing additional uncertainty into the estimation process.

Another possible limitation of the comparison of State performance to that of pseudo-States is that the standardization process described above adjusts only for the student characteristics found in the State and not for the extra consequences for the State of the concentrations of students with special needs. For instance, the nature of a State's efforts to respond to the educational needs of limited English proficient (LEP) students when such students represent a substantial proportion of the students in a significant number of schools within the State is likely to be qualitatively different from another State where LEP students are either less numerous or more widely dispersed. The sampling procedure described above cannot readily distinguish whether a case from a given cell of the national distribution was drawn originally from a school or State with higher low LEP concentration. (In technical language, the adjustment accounts only for the compositional influences of demographic characteristics but not for their contextual influences (e.g., Burstein, 1980).)

To the degree that complications already described can be minimized, the comparison of a State to its corresponding pseudo-States has intuitive appeal in that the procedure operationalizes comparability in terms of the State's own demography. Moreover, it does so in a manner which neither modifies nor masks the State's performance relative to the National-NAEP standard. For instance, one can readily envision a display that plots the performance levels of the States bounded within comparison bands based on the sampling distributions of their corresponding pseudo-States. A hypothetical example of this strategy might lead to a finding that Wisconsin's performance level of 275 is above the national average of 250 but falls at the 40th percentile of its distribution of pseudo-State samples while Tennessee's performance level of 245 is below the national average but falls at the 75th percentile of its distribution of pseudo-State samples. Both sets of results highlight achievements and target needs for improvement.

State to Itself Over Time

Just as monitoring achievement trends is the main reason for conducting National-NAEP, we anticipate that the ability to monitor State-level trends will eventually be seen as the primary benefit of State-level NAEP. The notion that States should be expected to progress over time seems inherently less controversial than attempting to assess relative status at a given point in time. Moreover, the benefits of monitoring progress separately by demographically defined subgroups and by important learning sub-domains, along the lines discussed in the CCSSO report (1988), are clear. Otherwise, there is a risk that aggregated trends in progress will mask disparities in the pervasiveness of progress. For example, a concentration on trends in total performance in the recent NAEP mathematics report (NAEP, 1988) could lead to misleading conclusions since the relative gains over time in lower-level mathematics skills, especially for blacks, masked the lack of progress on higher-level skills and applications for all subgroups over time. Similar findings at the State level might lead to policy decisions either to stay the present course or to alter the emphasis and targets of reform efforts.

Attempts to interpret trends at the State level are not without complications, however. Under current plans, State-level samples are likely to be relatively small (2,000-3,000 students per age/grade level) and thus State-level performance trends will be much less stable than national trends, especially at the level of demographic substrata. Moreover, any changes over time in either the State's demography or the quality and characteristics of State-level samples will affect the interpretability of trend data in much the same way as performance patterns on college admissions tests are influenced by changes in State-level participation rates in such testing programs. None of these complications are reason enough to exclude State-level trend reporting. Rather, they justify careful monitoring and reporting of trends in the characteristics of State samples in conjunction with achievement trends.

State to Absolute Scale

In recent NAEP reports (1985, 1988), there have been attempts to report results on scales that are anchored by student performance expressed in terms of the types of test questions that most students attaining a given score level would be able to answer correctly. This so-called proficiency scaling establisnes a correspondence between specific tasks and the underlying scale on which student performance is represented by the IRT (item response theoretic) methodology used by NAEP. Once established, the intent of proficiency scales is to provide a form of absolute standard against which to measure performance in a given assessment area. For example, one might judge that 60 percent of the sample of 17-year-olds can handle the mathematical tasks involving sophisticated numerical reasoning (a proficiency level of 300 according to the 1986 NAEP results) in 1990 where roughly 51 percent could do so in 1986.

Here, again, an informative method of comparison for the Nation can also be useful for State-level data. Monitoring a State's results at different proficiency levels better characterizes the distribution of the performance of its students and the nature of its strengths and weaknesses. For example, reporting that 95 percent of North Dakota's 13-year-olds can perform basic operations and beginning problem-solving but only 10 percent can handle sophisticated numerical reasoning tasks focuses the State's improvement efforts on those aspects of the curriculum where progress needs to be made.

One feature of the National-NAEP reporting of proficiency levels might need reconsideration before application to State data. Currently, NAEP proficiency scales span all age/grade levels participating in NAEP assessment. Thus the 1986 NAEP mathematics results indicate that roughly 20 percent of 9-year-olds, 73 percent of 13-year-olds, and 96 percent of 17-year-olds can perform basic mathematical operations and are beginning to develop problem-solving skills. While national trends along these lines may have some merit, targeting proficiency reporting to specific age/grade levels is likely to be more helpful at the State level. That is, the anchoring of proficiency scales to expectations about performance for a given age/grade level and how well each State's students perform with regard to such scales (as in the North Dakota example in the previous paragraph) would be more beneficial.

State to State

Much of the discussion in various reports and publications envision State-to-State comparisons. The Wall Chart has consistently reported such comparisons and the reports released by the National Governors Association (NGA, 1987) and CCSSO (1984, 1987, 1988) anticipate such comparisons. None of these efforts make the attempt to compare States directly any less hazardous and tricky. No two States (or other educational units, for that matter) are exactly comparable in either their student and schooling characteristics or in their educative intents, resources, and efforts.

Nonetheless, State-to-State comparisons will be made. Under such circumstances, it is important that the basis for such comparisons be as clear and as fair as possible. Conditions that encourage clarity and fairness are those spelled out in the principles articulated earlier. To the extent that multiply reported disaggregated data can be provided in a form that reflects the pertinent dimensions that might distinguish performance within any given State across the States, State-to-State similarities and distinctions in educational accomplishments are worth reporting.

The CCSSO report (1988) and Haertel's paper in this report each discuss several options for reporting State-to-State comparisons. The three options proposed by CCSSO include (a) comparisons to other States within regional clusterings like those used by NGA (1987); (b) comparison across States for students with similar background characteristics; and (c) comparison of achievement of States ranked according to a composite of State demographic characteristics. The options Haertel considers viable are (d) reporting State achievement without any kind of adjustment or clustering (with comparisons of performance levels of like students across States); (e) deriving State achievement comparison bands from first applying regression models to data units defined by assigning State sample data to community type within major geographic area and then aggregating predicted results to obtain State-level estimates; and (f) creating floating comparison groups of States defined by locating a State within clusters formed by choosing those States just above and below the State on a demographically defined continuum.

Once State-level NAEP data become available, it is likely that all of the options mentioned will be applied by some subset of users. If historical precedent holds, the media will portray raw performance rankings in some form of league table while NGA and CCSSO will most assuredly present results by geographically defined clusters of States.

One can also anticipate that comparisons based on demographically defined groupings will spark the most controversy. The controversy will be generated by differences of opinion about the types of demographic factors to take into consideration, whether to treat each factor separately or form composites, and once factors are selected, whether to cluster States or simply list them in ranked order. The CCSSO report (1988) contains an illustrative display with States ordered according to a composite of three background characteristics representing, respectively, the State's wealth (e.g., per capita income), the educational level of its citizenry (e.g., percent of adults who have completed 4 years of high school), and the poverty concentration within the school-age population (e.g., percent of school-age children who live in poverty). While these conditions are beyond the control of the educational system and thus reflect realistic constraints on a State's ability to develop and maintain a sound economic and social foundation for its educational institutions, both the means of choosing suitable measures of each condition (Cf., e.g., CCSSO, 1988, pp. 16-17) and of using the measures or a composite based on them are far from straightforward. For instance, States such as Utah and Wyoming with relatively high educational levels and low poverty concentrations also tend

to have low income levels, while States such as New York and Illinois combine high income levels with larger poverty concentrations and lower educational levels (which mask considerable intra-State heterogeneity). Or, alternatively, is California more like Maryland (its closest "neighbor" on the per-capita income figures reported by CCSSO), Kansas (educational level), Illinois (poverty concentration), or Oregon (the unweighted composite of the three background indexes)? Obviously, there will be debates about which means of characterizing background conditions best reflects the intent to take educative difficulty into consideration.

Concluding Comments

We have attempted to portray the issues, the options, and the complications inherent in reporting State-level NAEP performance in a fair and creditable manner. While we argued that certain principles should guide intent, most notably those dealing with disaggregation of results wherever substantively warranted and technically feasible and sensible, straightforward, uncontroversial application of the principles under the anticipated conditions of the two State-level trials is unlikely.

Concerns about the difficulties in an a priori resolution of differences of opinion about the best means of reporting State-level data and State comparisons need not be debilitating. The panel recommended that a wide variety of alternative methods be employed during the 1990 and 1992 trials. Our sense is that it will be important to put forward as many systems of reporting State-level results as the NAEP contractor, the NAEP governing board, the National Center for Education Statistics, NGA, CCSSO, the media, and other interested parties can devise during the trial period.

The empirical evidence from the trials should provoke discussion and debate about the relative merits of different reporting systems. Out of such debates could evolve a set of methods that reflect, either implicitly or explicitly, a working consensus among the various constituencies and thus become the core reporting methods for State-NAEP after 1994. On the other hand, if a consensus is unachievable through such a process, the reports produced by the different constituencies are likely to reflect their distinctive institutional and organization frameworks and perspectives. Neither consequence represents either an uncommon or unhealthy situation. Both conditions could encourage attention to the similarities and differences within and across States in the nature of their educational circumstances and accomplishments. If so, they can foster ongoing dialogue about the health of the Nation's educational system, which was the implicit intent of calls for State comparisons in the first place.

References

Alexander, L. & James, H. T. (1987). The Nation's report card: Improving the assessment of student achievement. Report of the Study Group. Cambridge, MA: National Academy of Education.

Averch, H. A., Carroll, S. J., Donaldson, T. S., Kiesling, H. J., & Pincus, J.(1972). How effective is schooling? A critical review of research. Santa Monica, CA: RAND Corporation.

Brookover, W. B., Beady, C., Flood, P., Schweitzer, J., & Wisenbaker, J. (1979). School social systems and student achievement: Schools can make a difference. New York: Praeger.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D. C. Berliner (ed.), Review of research in education, Volume 8. Washington, D.C.: American Educational Research Association, 158-233.

Burstein, L., Baker, E. L., Aschbacher, P., & Keesling, J. W. (1985). Using state test data for national indicators of education quality: A feasibility study. Los Angeles: Center for the Study of Evaluation, Graduate School of Education, UCLA.

California State Department of Education (1977). School effectiveness study: The first year. Sacramento, CA: Office of Program Evaluation and Research.

California State Department of Education (1986). Performance report for California schools. Sacramento, CA.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F. & York, R. (1966). Equality of educational opportunity. Washington, D.C.: U. S. Government Printing Office.

Council of Chief State School Officers (1984). Educational assessment and evaluation in the United States. Position paper, Washington, D.C.

Council of Chief State School Officers (1987). Volume I: State Education Indicators 1987. Washington, D.C.: State Education Assessment Center.

Council of Chief State School Officers (1988). On reporting student achievement at the State level by the National Assessment of Educational Progress. Recommendations from the National Assessment Planning Project. Washington, D.C.

Dyer, H., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 6, 591-605.

Edmonds, R. R. (1979). Effective schools for the urban poor. Educational Leadership, 37, 15-27.

Klitgaard, R. E., & Hall, G. R. (1974). Are there unusually effective schools? Journal of Human Resources, 74, 90-106.

Mosteller, F. & Moynihan, D. P. (Eds., 1972). On equality of educational opportunity. New York: Vintage Books.

Purkey, S. C. & Smith, M. S. (1983). Effective schools: A review. Elementary School Journal, 83(4), 427-452.

National Assessment of Educational Progress (1985). The reading report card: Progress toward excellence in our schools, trends in reading over four national assessments, 1971-1984. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress (1988). The mathematics report card: Are we measuring up? Trends and achievement based on the 1986 national assessment. Princeton, NJ: Educational Testing Service.

National Governors Association (1987). Results in education. Washington, D.C.

New York State Department of Education (1976). Three strategies for studying the effects of school process. Albany, NY: Bureau of School Programs Evaluation.

Selden, R. W. (1986). White Paper: Strategies and issues in the development of comparable indicators for measuring student achievement. Washington, D.C.: State Education Assessment Center, CCSSO.

Southern Regional Education Board (1987). Measuring student achievement: Comparable test results for SREB States and the Nation. Atlanta, GA.

U.S. Department of Education (1984). State education statistics: State performances, resource inputs, and population characteristics 1972 and 1982. Washington, D.C.

Within-State Comparisons: Suitability of State Models for National Comparisons

Edward Haertel
ASA Fellow, NCES

Student achievement tests have assumed unprecedented importance as indicators of educational outcomes and as tools of educational policy. In a decade marked by concern over educational accountability, test scores have stood alone as obvious, objective, and available indicators of significant schooling outcomes. The SAT test score decline, references to test performance in A Nation at Risk and other reform reports, and the annual Department of Education "wall charts" ranking the States on achievement and other education indicators have all contributed to a heightened interest in test results.

Test scores assume meaning largely through processes of comparison. A raw score on an achievement test assumes meaning when it is expressed as a percentile or grade equivalent, placing it in the context of test scores for some meaningful comparison group. Likewise, an average SAT score of 471 means little in itself, but acquires meaning when it is expressed as an improvement or a decline over the previous year's performance. Finally, average test scores for schools, school districts, or States may take on additional meaning when they are compared to those for other schools, districts, or States. This last form of comparison is the topic of this paper.

In this paper, several methods are described for presenting comparisons among schools or districts within a State. Each is illustrated with descriptions of the specific procedures used in one or more States. The paper concludes with a discussion of the suitability of these methods for making comparisons among States within the Nation. The first method described, raw comparison, is simply to report and rank unadjusted school and district means, as is done in Iowa. Next, methods of reporting achievement relative to a range of expected scores for each school are described, and illustrated with the procedures used in Pennsylvania. Three more approaches each limit comparisons to more-or-less homogeneous subsets of schools or districts. Stratification methods, illustrated by the system used in New Jersey, rank all districts according to an index of the educational challenge their students present (essentially an index of socioeconomic status), then stratifies the districts according to that ranking, and makes comparisons within strata. The procedure used in Massachusetts is to cluster districts according to community type, then make comparisons within clusters. Finally, the method of floating comparison groups used in California is described. Under this method, schools are ranked and each is then compared to its own unique comparison group comprising some fixed number of schools ranked above and below it.

Before turning to these various methods of reporting, some assumptions underlying any comparative reporting method are briefly discussed. Following the presentation of the different methods, the paper concludes with a general discussion of the implications for State-by-State comparisons.

Valid Comparisons Among Educational Units

Suppose a school district includes four elementary schools. Schools A and B serve largely "at risk" students, and schools C and D serve largely "advantaged" students. A and B have lower average daily attendance, higher transiency, and more LEP students. They serve areas of the city with higher unemployment, and on average, the parents of students in schools A and B have fewer years of education than parents of students in C and D. There are also fewer books and magazines in their homes. A greater proportion of students in schools A and B come from single-parent families.

Suppose now that the district selects a standardized test that validly measures some of the learning outcomes in its curriculum, and administers that test to all fourth graders in the four schools, following proper testing procedures. Suppose further that the average test scores for schools A and B turn out markedly lower than for schools C and D. Are comparisons among these average scores for the four schools valid? The answer can only be, "It depends." Such comparisons are valid for some purposes, not for others. It is probably correct to infer that fourth graders in schools A and B are not performing as well as those in C and D on the objectives measured. It is not correct to infer that schools A and B are less well run or have less effective teachers. Achievement comparisons alone cannot answer questions about school quality when the schools compared serve different kinds of students.

Test score comparisons among schools, districts, or States invite inferences about the relative quality of services those schools or systems provide. Indeed, the logic of testing for accountability all but demands such inferences. Differences among the units[1] compared in the educative environments of students' homes and communities, in student language backgrounds, or other factors render suspect simple comparisons among average scores. This suggests that reporting systems should incorporate some kind of adjustment for such differences. An ideal reporting method would indicate which units were doing well and which were doing poorly relative to the levels of achievement that ought to be expected of them.

Risk of legitimating inequality of educational outcomes. The preceding discussion may have suggested that making fair comparisons among test score averages was basically a technical problem. Quantifying the differences in the amount of challenge posed by the student bodies in different schools or school systems might be difficult, but still in principle a purely statistical matter. In fact, of course, the definition of fairness in comparisons is bound up with educational philosophy and values. The challenge set was to indicate units' performance relative to the levels of achievement that ought to be expected of them. One position holds that in our society, the same expectations for content mastery must be held for all

students. To judge an inner-city school satisfactory for doing better than other inner-city schools is unacceptable if its students still fall short of national averages. It follows from this position that any methodology that adjusts for background characteristics runs the risk of legitimating existing inequities by implying that inferior outcomes are good enough for students that have historically achieved at inferior levels. Thus, there is a tension between invidious comparison and legitimation of inequality. On the one hand, it seems unfair to hold schools or systems to a common expectation when they enjoy unequal levels of out-of-school support. On the other hand, it seems unfair to accept unequal outcomes for different learners, regardless of the rationale.

This is a serious and abiding problem, not to be resolved in this paper. The methods to be discussed for making test score comparisons all are referenced, implicitly or explicitly, to the status quo. In one way or another, all of the methods employed by the different States embody an assumption that on average, schools serving different kinds of students are about equally good. Equivalently, these methodologies assume that the average differences in achievement across levels of socioeconomic status (SES), size and type of community (STOC), or other factors adjusted for are due entirely to differences in the educative challenge posed by different kinds of students, rather than differences in the average quality of education offered by their respective schools or school systems.

Consider a possibly hypothetical example. If large-city schools in general offered poorer educational services and in general served less advantaged students, then the gap between their achievement scores and those of smaller cities would be due in part to the greater challenge their students posed and in part to the poorer instruction delivered. (These effects would not necessarily be additive--if poorer students were more vulnerable to the effects of poor instruction, then the interaction between student and school characteristics would further increase the achievement disparity.) One way of making "fair" comparisons among schools would be to cluster them according to size and type of community, and then compare each school to others within its own cluster. This approach would indicate that typical large-city schools were doing about as well as could be expected. Another way of making "fair" comparisons would be to regress average school achievement on some SES composite (also defined at the school level), and to use the predictive equation obtained to calculate an expected achievement level for each school. This approach would also indicate that typical schools serving low-SES students were doing about as well as could be expected. Neither approach would reveal overall differences in the quality of schooling provided to large-city versus smaller-city students, or to high-SES versus low-SES students. Differences due to instructional quality would in effect be attributed to student demographics.

Appropriateness of the educational objectives tested. One last assumption must be mentioned in passing. Fair comparison also requires a test that validly measures learning outcomes given the same priority in all of the units compared. If the schools within a district or the districts within a State are supposed to teach the same curriculum, then a test chosen to represent that curriculum should satisfy this assumption. For

interstate comparisons, however, finding a test that offers a fair basis for comparison might be more problematical. If a given learning outcome is not equally weighted in the curriculum frameworks of different States, then it might be unfair to compare those States on a test of that learning outcome. Note that in testing for educational accountability, the learning outcomes covered should have equal priority in the intended curricula of the schools or systems compared. Equal representation of these outcomes in the instructional materials used or in the instruction delivered does not bear on the question of accountability.

Summary. Despite their limitations, risks, and hidden assumptions, methods for comparing schools or school systems can be useful in guiding educational policy. Expected achievement levels or observed achievement in similar units can be used to set realistic goals for improvement, to recognize excellence, and to target resources to the areas of greatest need. Such methods must never be used, however, to legitimate inequalities in educational outcomes. Raw, unadjusted achievement scores also say something important about the relative attainments of students, and should always be reported in conjunction with any adjusted or expected scores. If comparisons are made within cluste of schools or systems that resemble one another, performance should also be reported relative to schools or systems across all clusters. In fact, it appears to be universal practice to present unadjusted comparisons to the entire set of schools compared in conjunction with comparisons to any adjusted or predicted scores.

Models for Comparing Districts or Schools

Reporting Unadjusted (Raw) Achievement Scores

Iowa has the oldest and most comprehensive testing program of any State, although it is not administered by the State government. Both public and private schools participate. For decades, virtually every elementary school student at every grade level has taken the Iowa Tests of Basic Skills (ITBS) in the fall of every year. Over 80 percent of the high schools in the State administer the Iowa Tests of Educational Development (ITED) annually, and most of the remaining high schools give the ITED every other year. At all levels, the reporting of individual performance to pupils, teachers, and parents is emphasized, but information about schools and districts is also prepared and made available. In addition to student-level norms (e.g., percentile ranks for individual students' scores), school-level norms are prepared,[2] and the roughly 430 districts in the State are ranked from highest to lowest.

Simply reporting the relative standings of students, of schools, and of districts has some obvious advantages. It is easy to understand, and it avoids entirely the dangers of legitimating unequal outcomes by setting different expectations for different learners. Its major disadvantage is the difficulty of reaching judgments about the relative quality or effectiveness of different schools or districts, which may serve different sorts of student populations. The system works in Iowa primarily because no great weight is placed on it. School districts are legally required to

release their test results to whomever requests them, but school-by-school reporting of test results in local papers is unusual. There are no fiscal or other rewards or sanctions associated with good or poor performance for schools or districts. The primary emphasis is on individual-level reporting of performance to students, teachers, and parents. Another contributing factor may be the relative homogeneity of students, schools, and districts in Iowa. Many States embrace far greater extremes of educative challenge and of schooling outcomes.

In summary, Iowa's testing program highlights the importance of the intended uses of test score comparisons in any consideration of ways in which those comparisons should be made and reported. The need for elaborate adjustments or comparison procedures only arises when considerable public attention is directed to test score comparisons, or when fiscal or other rewards or sanctions depend on them.

Expected Scores

It is a universal finding that student achievement test scores are correlated with socioeconomic status. Although the full range of student achievement levels may be found at all socioeconomic strata, it remains true that on average, students whose parents earn more money and have completed more years of schooling will themselves earn higher scores on achievement tests. Suppose that one wished to compare schools serving student bodies at different socioeconomic levels. Assuming that each school's SES could be quantified, regression analysis could be used to adjust for these differences.

The predicted achievement levels provided by a regression analysis are no more or less than conditional means--average levels of achievement among those units at some given level of SES. Linear regression obtains these conditional means under the assumption that there is a linear relationship between mean SES and mean achievement at the school level. If rather than some single SES composite, measurements of several variables are used, then multiple linear regression can yield the weighted sum of those measurements that best predicts average achievement. A school's observed average achievement score is compared to its predicted score, derived via multiple linear regression. If various assumptions of the regression model are satisfied, then this is equivalent to comparing the school to the average achievement in a hypothetical population of schools having its exact profile of background characteristics.

Not only the mean, but also the variance of school achievement levels may be estimated for the school's hypothetical comparison group. A "comparison band" may be reported, including a range of some number of standard deviation units (standard errors) above and below the hypothetical mean. A school whose observed achievement falls within this range is performing "as expected," and those above or below their comparison bands may be singled out as exceptional. The width of the comparison bands determines about what fraction of all the schools in the State will be designated high or low achieving relative to "as expected."

This kind of regression adjustment to achievement test scores is ubiquitous in educational research. The original "effective schools" research began with a search for statistical outliers, that is, schools achieving at substantially higher levels than predicted by the characteristics of the students and communities they served. Some standardized achievement tests (e.g., the Metropolitan Achievement Tests) offer among their standard scoring services an optional report comparing each school building's mean performance to the level predicted from an index of teacher- or student-reported parent educational level.

In Pennsylvania's Educational Quality Assessment (EQA) program, achievement scores are obtained for each school at each grade level, for each of a series of content areas. These school scores are averages of the scores of all students participating in the (matrix-sampled) assessment.[3] An equation is then derived to predict these school means using the variables shown in Table 1. The same variables are used across grade levels and content areas. They were selected from among available "nonmodifiable" or out-of-school background variables representing student socioeconomic level, based in part on findings from stepwise regression analyses, and are intended to represent influences on achievement over which schools have no control. The inclusion of a variable does not imply that it has some direct or causal influence on achievement, merely that it serves as a proxy for some complex of background factors related to achievement (Pennsylvania Department of Education, 1987).

Table 1.--Variables Used for Prediction of School Mean Achievement Scores in Pennsylvania's EQA Program

Predictor	Source of Data
Percentage of low-income students	State Chapter 1 files
Percentage of girls	Student questionnaire (self report)
Level of parental education	Student questionnaire
Population density of residential community[1]	Student questionnaire (assisted by examiner if necessary)
Percentage of white students	Student questionnaire
Frequency of residence/school change	Student questionnaire
Student time spent watching television	Student questionnaire
Number of books and magazines in the home	Student questionnaire

[1]The square of this variable is also included in the regression equation.

Prediction bands are obtained as each predicted score plus or minus one standard error. Under this formula, about two-thirds of all school scores would be expected to fall within their comparison bands, about one-sixth would fall below their bands, and about one-sixth would fall above. Note that the choice of one standard error as the width of the comparison band is arbitrary. It is effectively a decision to flag the lowest and the highest 16 percent of the adjusted scores, rather than, say, the lowest and highest 10 percent or the lowest and highest 25 percent. Note also that this procedure does not account for differences in the precision of achievement averages according to the number of students tested. A slightly more complex formula would provide narrower bands for larger schools than for smaller schools.

Each school receives a report showing, for each grade level/content area combination, the number of students on which the score is based, the obtained school mean, its percentile rank and stanine in the overall State distribution, and an indication of whether it is below, within, or above the calculated comparison band.

If the assumptions of the regression model are satisfied, it yields comparisons of schools "holding constant" the effects of whatever variables are included in the equation. Thus, the choice of these variables is critical. In the Pennsylvania procedures, it is emphasized that predictors are limited to "nonmodifiable" or out-of-school variables, on the grounds that these represent the "inputs" or "raw materials" the schools have to work with (Pennsylvania Department of Education, 1987). Ideally, adjustments for unmodifiable variables would account for the particular characteristics that made a school's student body different than average. If the school's actual achievement was above this predicted level, it would follow that it was a superior school, and conversely.

This rationale for including only "nonmodifiable" variables in the equation may be clarified by considering a counterexample. Suppose that a "modifiable" or school process variable were included--amount of homework, for example. Each school would then be compared to a hypothetical population of schools that among other things assigned as much homework as it did. Schools that assigned too little homework would not be penalized, because their hypothetical comparison groups would consist of other schools that assigned the same amount of homework that they did. Thus, there would be no incentive for increasing homework.

The <u>logic</u> of adjusting only for unmodifiable variables is compelling, but unfortunately, the practice is not so simple. A complete discussion of the assumptions entailed, including the measurement (without error) of <u>all</u> such variables and the linearity of their relationships to achievement, may be found in standard texts on regression. The most critical assumption for purposes of this discussion is that the set of unmodifiable variables did <u>not</u> predict any of the variance in school process variables, which of course do not appear in the equation. This is a more formal statement of the concern raised initially that calculating and reporting expected scores may legitimate inequities in educational services by attributing achievement disparities entirely to differences in educational inputs. To

continue the earlier example, if the amount of time students spend doing homework could be predicted in part from the amount of time they spend watching television or from the percentage of low-income students, then including these variables in the predictive equation would have some of the same effect as including amount of homework assigned. Limiting the predictors included in the regression equation to those that appear to measure "out-of-school" factors does not assure that only "inputs" or "raw materials" are being adjusted for. Not only may these procedures fail to adjust for some part of the out-of-school effects, but they may also erroneously adjust for some effects of school policies and practices.

The commentary provided by the Pennsylvania Department of Education (1987) includes firm, clear, and appropriate warnings against assuming that the variables in its equations are the only important ones, inferring cause-and-effect relationships from the regression equations, or inferring the relative importance of the background variables from their ordering in the regression equation or from the relative magnitudes of their regression coefficients. Similar warnings should be issued any time these procedures are applied.

In spite of the limitations of regression procedures for estimating expected or predicted scores, they are widely used and well understood. The Pennsylvania procedures described by way of illustration appear to represent a technically sound, sensitive, and appropriate application of predicted scores as an adjunct to the reporting of unadjusted comparisons to State means.

The three remaining methods resemble one another more closely than they resemble the approach used in Pennsylvania. Expected score methods like Pennsylvania's use data from all schools or districts to estimate a model relating achievement levels to out-of-school characteristics, then use that model to produce expected achievement levels for each of the units compared. The results are comparisons of observed achievement levels to mathematical predictions of what those achievement levels should be. The three remaining methods all present comparisons of each school or district to other, actual schools or districts, not to predicted scores. These other schools or districts are chosen to be a fairer comparison group than the set of all schools or districts would be. The three methods differ primarily in how these other schools or districts are selected.

Stratification Methods

One simple way to compare a school or district to others like itself is to arrange all of the units along some continuum from lowest to highest, divide that continuum into several strata, and make comparisons within strata. The system used in New Jersey for organizing school districts into District Factor Groups (DFGs) will serve as an illustration. Procedures of this kind require three steps: (1) defining the continuum; (2) creating the strata; and (3) expressing each school's or district's performance relative to that of others in its stratum. These are discussed in turn.

Defining the continuum. Schools or school systems might be ordered according to any number of different variables, with different choices giving different rankings. Different rankings would favor different districts, and so the method chosen must be carefully considered and well justified. In many ways, the problem of choosing a continuum to define equitable comparison groups resembles that of choosing a set of variables to calculate expected scores. Districts located close together on the scale should resemble one another with respect to characteristics that are strongly related to achievement but beyond the control of the district. Many different factors might be chosen, including home educative environment, parent education, student transiency, or economic factors related to out-of-school learning opportunities. Each of these (and of course other possible variables) would define a somewhat different continuum. However, most such variables are positively correlated with one another, and so a composite of several such indicators is likely to correspond fairly well to any one of them, including those not directly represented in the composite. The continuum so defined will turn out once again to represent something like socioeconomic level.

Given district achievement score averages and a set of background variables, a continuum could be defined by determining what weighted combination of background variables best predicted the achievement scores. This could be done using multiple regression, exactly the way the expected achievement scores were calculated for schools in Pennsylvania. The background variables would be used to calculate predicted achievement levels for each district and these predicted achievement levels would define the continuum. Districts would then be ordered according to their predicted achievement levels, so that those located close to one another on the continuum had sets of background characteristics yielding similar predicted achievement levels. Note that even though achievement data were used in determining the regression equation defining such a continuum, each district's ranking would be solely a function of its background characteristics, and not of its achievement scores.

Rather than explicitly using information about the relation of different background variables to achievement, alternative procedures might seek a composite that is most effective in distinguishing among districts. In other words, one might seek not the weighted combination of background variables that correlates most highly with achievement, but rather, the weighted combination that maximizes variability from one district to another. The statistical procedure appropriate for this approach is principal component analysis, which was used to define the DFGs presently employed in New Jersey.

The background variables used to define New Jersey's DFGs are shown in Table 2. Information on these variables was obtained from the 1980 census for each of the roughly 500 school districts in New Jersey. The first principal component of the covariance matrix of these variables was used to define the continuum according to which districts were ranked. New Jersey's DFGs date back to 1974, and procedures for defining the continuum have been modified from time to time. Initially, factor analysis rather than principal component analysis was employed, and an eighth variable,

mobility, was included. Mobility was defined by the percentage of persons residing in the same housing unit for at least the past 10 years. It was eliminated from the equation based on empirical findings that it did not contribute significantly to the differentiation of districts according to socioeconomic status.

Table 2.--Variables Used to Define District Factor Groups (DFGs) in New Jersey

Educational Level	
1	= less than 4 years of high school
. . .	
5	= more than 4 years of college
Occupational Status	
1	= laborers
. . .	
11	= old and new professionals
Density	
	number of persons per household
Urbanization	
	percent of district considered urban
Income	
	median family income
Unemployment	
	Percent of those in the work force who received some unemployment compensation in 1979
Poverty	
	Percent of residents below the poverty level in 1979

Note: All data are obtained from the 1980 decennial census.

Although factor analysis could be informative concerning the underlying structure and dimensionality of the district background variables, use of component analysis to define the SES composite is more appropriate. Estimating the score of each district on a hypothetical factor would be much more complex and controversial than calculating its score on the first

principal component. Component analysis is the proper statistical tool to use.

Creating the strata. Having arrayed the districts along a continuum, they must then be divided into strata. This might be done by dividing the continuum into equal intervals, or by dividing the ranked districts into equal-sized groups. The latter approach was actually taken, creating ten DFGs of about 50 districts each. An eleventh DFG was created for vocational districts, which draw their students from larger geographical areas. There appears to be no compelling reason for requiring that the DFGs be of uniform size. Suppose that the distribution of district SES is more or less bell shaped, with most districts at intermediate levels and fewer in the extremes. Then the lowest or highest decile of the districts would span a wider range of SES levels than an intermediate decile, and so within-DFG heterogeneity would be greater in the extreme groups. Rather than forming groups containing equal numbers of districts, it would be possible to construct groups spanning approximately equal ranges of SES. This alternative approach of dividing the SES continuum into equal intervals would reduce the problem of differences in within-DFG heterogeneity, but could easily result in some DFGs containing only a few districts. Moreover, the equal-interval solution would be strictly justified if there were a linear relationship between the SES scale and the district's average expected achievement. The procedure followed, dividing the districts into deciles, depends only on their rank ordering, and does not involve any assumption that the SES continuum is an equal interval scale, or that it is linearly related to achievement.

By design, each DFG is substantially more homogeneous than the set of all 500 or so districts in New Jersey, and so each district resembles others in its own DFG more closely than those in other DFGs. Nonetheless, each DFG still defines a range of socioeconomic levels. Districts near the boundaries between DFGs may be slightly penalized if they fall at the bottom of a group of higher SES districts or slightly favored if they fall at the top of a group with lower SES districts. Increasing the number of DFGs reduces this problem, but makes each DFG smaller and therefore less stable. The decision in New Jersey to create comparison groups of about 50 districts represents a compromise between homogeneity and stability.

Describing each district's performance. The DFG serves as a norm group for each district it contains. In addition to reporting a district's performance relative to all districts in the State, its achievement score averages may be reported relative to the distribution of achievement averages for its DFG. The district's quartile, percentile rank, or stanine within its DFG might be reported, for example.

Clustering Methods

As part of its complex system for reporting school achievement test results, Massachusetts uses a community classification scheme developed in 1985. Unlike New Jersey's DFGs, Massachusetts's Kind of Community (KOC) categories were not created by segmenting an SES continuum, nor can they be arrayed along a single dimension. Fifteen socioeconomic and demographic

variables were identified using data from the 1980 census and from State agencies, including variables reflecting community property values, income, educational level, economic activity, percent minority, population density, and other factors. After obtaining information on these 15 variables for each of the roughly 350 communities in the State, cluster analysis was used to find homogeneous community categories. The KOC categories that resulted are shown in Table 3 (Massachusetts Department of Education, 1985).

Table 3.--Kind of Community (KOC) Categories Used by the Massachusetts Department of Education

Urbanized Centers	Manufacturing and commercial centers; densely populated; culturally diverse
Economically Developed Suburbs	Suburbs with high levels of economic activity; social complexity; and relatively high income levels
Growth Communities	Rapidly expanding communities in transition
Residential Suburbs	Affluent communities with low levels of economic activity
Rural Economic Centers	Historic manufacturing and commercial communities; moderate levels of economic activity
Small Rural Communities	Small towns; sparsely populated; economically undeveloped
Resort/Retirement and Artistic	Communities with high property values; relatively low income levels, and enclaves of retirees, artists, vacationers and academicians

Source: A New Classification Scheme for Communities in Massachusetts. Massachusetts Department of Education, 1985. (Publication No. 14253-1500-11-85-CR)

Although factors like community educational level, unemployment rate, and percentage minority are reflected in the KOC categories, these categories are probably not sufficiently homogeneous to define comparison groups for schools. Accordingly, classification by KOC is just the first step in Massachusetts's reporting system. Within each separate KOC, regression analyses are used to predict achievement using four background factors that reflect a school's socioeconomic status. Separate regressions are done for each grade tested (3, 7, and 11) in each broad content area (reading, mathematics, and science). The background factors are obtained

are done for each grade tested (3, 7, and 11) in each broad content area (reading, mathematics, and science). The background factors are obtained from either a school principal's report (grade 3) or student questionnaires administered as part of the State assessment (grades 7 and 11). At grade 3, they include the school's average parent occupational level (calculated from the principal's report of the proportions of parents in each of five categories), percent receiving free or reduced price lunches, percent who left school since the beginning of the year, and percent whose families' native language is not English. At grades 7 and 11, the background factors included indices of mothers' and of fathers' education, of language other than English in the home, and of the proportion of students born outside the United States (Massachusetts Department of Education, 1986).

Regressions of overall reading, mathematics, and science scores on these background factors yield predicted achievement levels for each school, and these become the midpoints of comparison bands for each school. The width of the comparison band is determined so that in a given content area/grade level combination, 25 percent of the schools fall below their bands, 50 percent within the bands, and 25 percent above the bands. These band widths account for both the standard error of estimate and the number of students tested, using a formula of the form

(constant 1) + (constant 2) / (square root of number of test scores used)

where (constant 1) and (constant 2) depend on the grade level and content area.

Like the other States discussed, Massachusetts provides raw comparisons to the entire State distribution, as well as comparisons adjusting for socioeconomic level. In its Educational Assessment Report, columns appear for the entire State, for the school's KOC category, for the school's district, and for the school itself. These are followed by a column giving the school's comparison band. There is a row for each test.

Floating Comparison Groups Method

The State of California has developed a variant of a stratification method that merits special attention. It involves the same three steps of defining a continuum, creating strata, and expressing each school's performance relative to that of others in its stratum. With the floating comparison groups method, however, every school between the tenth and ninetieth percentiles of the SES continuum is located at the midpoint of its own uniquely defined stratum. Thus, it is in the second step that the floating comparison groups method differs from the stratification method.

Defining the continuum. Four background factors are defined for each school, as shown in Table 4. Most of this background information is obtained from a student questionnaire administered in conjunction with the State assessment, with responses averaged across students within a school. Achievement scores for each school are obtained from the matrix-sample State assessment using item response theory (IRT) methods, in each of reading and mathematics. To locate the schools on a socioeconomic status

continuum, the reading and mathematics scaled scores are added together, and this sum is regressed on the four background factors. This provides an equation giving each school's expected achievement score as a function of its background characteristics. Schools are ranked according to these expected scores. Note that this regression approach was discussed earlier as a possible alternative to the procedure used in New Jersey. Note also that each school's location on the SES continuum is determined without any reference to its achievement scores, strictly as a function of the school's background characteristics.

Table 4.--Background Variables Used to Define SES Composite for California

Educational Level Obtained by More Educated Parent
1 = not a high school graduate
. . .
5 = advanced degree
Student Mobility
percent of students enrolled in the district during the last two years
English Language Fluency
percent Limited English Proficient (LEP) according to State criteria
Poverty
percent of families receiving Aid for Families with Dependent Children (AFDC)

Creating Strata. Several years ago, California used five strata defined by the quintiles of the SES ranking just described. The strata so defined were satisfactory for schools that happened to fall near the center of their comparison groups, but were less than satisfactory for districts near the boundaries between quintiles. Not only were such districts relatively superior or inferior to their comparison groups depending on which side of the boundary the fell on, but an unacceptable proportion of such schools were reclassified from year to year, so that they would fall at the top of one band 1 year, and at the bottom of another band the next year. The solution found in California was to define a different comparison band for each school, consisting of the 10 percent of schools above it on the socioeconomic ranking and the 10 percent below. (For schools in the top or bottom 10 percent of the entire distribution, the comparison band is defined as it was before, as the top [bottom] 20 percent of all schools.)

Describing each school's performance. As for other States, each school's performance is reported first of all relative to the overall distribution of California schools. Also reported are each school's percentile ranks in the distributions of scores for its comparison group. If, for example, a school's reading score surpasses those of 35 percent of

the schools in its own comparison group, then that school's reading percentile is 35. Percentiles defined in this way are not so easily compared across schools as conventional percentiles would be, because they are referenced to different distributions. Nonetheless, an empirical examination by Fettler (1988) showed that these "floating percentiles" had surprisingly good distributional properties, and that they were highly correlated (r = .90) with residual scores derived by conventional regression procedures.[4]

Summary

Across the States, there are two broad approaches to the problem of accounting for socioeconomic differences in comparing achievement across schools or districts. One is to generate predicted scores for each unit compared, and the other is to provide a more appropriate and homogeneous comparison group for each unit, comprising other units that it resembles. These approaches are used as adjuncts, never replacements, for straightforward reporting of each unit's actual achievement relative to the State as a whole. One of the most sophisticated reporting systems examined, for the Commonwealth of Massachusetts, combines both approaches by adjusting for socioeconomic level within clusters defined by different kinds of communities.

Reporting distributions as well as means. These methods, and the discussion of them to this point, have been limited to predicting average achievement levels for entire units. More elaborated reports of score distributions and of performance for student subpopulations could be far more informative. As one moves from the level of schools or districts to larger aggregates like States, more differentiated reporting becomes more feasible just because sample sizes can be larger. Figure 1 illustrates the format used to present student score distributions in the Educational Assessment Report of the Massachusetts Educational Assessment Program. Similar formats are used in some other States, as well. In the Massachusetts report, the quartile breaks for the entire State are used to define four achievement ranges, and these ranges are used to present achievement distributions for successively narrower comparison groups for a given school: its kind of community (KOC) and its district. A separate table in the Massachusetts report presents mean test scores for the State, KOC, district, and school, together with the school's comparison score band, all for a series of content areas.

For purposes of making SES adjustments, the mean appears to be the most tractable distributional summary. Nonetheless, statistical models could in principle be devised for other distributional parameters, such as medians, quartiles or other quantiles. Given the substantial difficulties in modeling even means successfully however, and given that models for means seem to have satisfied the accountability and policy requirements of the States, there appears to be little reason to pursue the problem of accounting for effects of socioeconomic level on other distributional characteristics.

Even if predicted scores or comparison bands are created only for means and not for other statistics summarizing distributions, fuller reporting of achievement distributions remains an important goal. The Massachusetts report demonstrates that easily interpretable reporting formats can be devised to communicate information about score distributions and norms for distributions.

Figure 1.--Format used in Massachusetts for reporting school score distributions in comparison to State, comparison group, and local norms.

Student Score Distribution

Content Area	Quarter	State %	Kind of Community %	District N	District %	School N	School %
Reading	Highest	25					
	Third	25					
	Second	25					
	Lowest	25					
Mathematics	Highest	25					
	Third	25					
	Second	25					
	Lowest	25					
Science	Highest	25					
	Third	25					
	Second	25					
	Lowest	25					

A table following the above format is included in the Massachusetts Department of Education's "Educational Assessment Report." The accompanying text explains that each entry provides the number or percentage of students who scored in a particular quarter of the Statewide distribution.

Reporting performance for student subpopulations. In addition to more fully describing score distributions for the entire population, test performance might be reported for student subpopulations. Such more differentiated reporting could reduce the need for SES adjustments, by permitting cross-State comparisons of students more closely resembling one another. Suppose, for example, that State A has lower achievement than

State B in part due to A's higher proportion of low SES urban students. Rather than comparing overall achievement for A and B, it would be more informative to compare low SES urban students in A versus those in B, and similarly for other demographic classifications.

If socioeconomic level were included in the scheme for defining student subgroups, then in theory reporting by subgroup could make other forms of adjustment unnecessary. A possible disadvantage, however, is the substantially greater amount of testing that might be required to obtain accurate achievement score estimates for all subgroups, especially in smaller States.

Models for Comparing States

In using any of these school or district models for comparing States, the first problem encountered is that there are so few units to be compared. Regression models for predicting achievement levels work best with at least 200 or so units, not a mere 50. If States were first divided into clusters, say into four geographic regions, the number of units within each cluster would probably be too small to support any further modeling by State SES. Even if within-State models cannot be applied directly at the national level, however, they may offer useful points of departure in considering methods for State-level comparisons.

Reporting Achievement Without Adjusting or Clustering

Complex adjustments or comparison groups may not be necessary. One alternative is still to simply report achievement for each State, without attempting to specify how high each State ought to score. To be fair to each State, data on overall achievement would be supplemented by reports of achievement for student subgroups within each State, as described above. Data would be reported so as to encourage interstate comparisons of like students, and to minimize attention to comparisons of overall means. In order to control sufficiently for achievement differences, the student subgroups used in such a system would probably have to incorporate socioeconomic level. For example, they might be defined by a cross-classification of socioeconomic level and kind of community. This form of reporting scheme would offer useful data and permit fair comparisons, but would be expensive. Larger samples would be needed to estimate achievement for each subgroup than would be required only to estimate achievement for the State as a whole. This scheme would also require linking SES information to individual students, possibly by using student-reported parent occupation or education.

Calculating Expected Scores for States

The States are few in number, and nearly all encompass a wide range of types of communities and of socioeconomic levels. For these reasons, models that predict State achievement means directly from State background variable means appear unpromising. As an alternative, one general approach is to build models for units at some lower level of aggregation than the

States, and then combine predicted achievement scores for these within-State units to get State-level estimates. This may be referred to as a composition approach.

In outline, a composition model would first require definitions of some relatively small number of types of within-State units. For example, these might be several types of school districts, types of communities, or types of students. Second, each State would be partitioned into units of the designated kinds. Third, statistical models would be formulated for achievement for each kind of unit. At their simplest, such models would predict the same achievement level for all units of a given kind. More sophisticated models might predict achievement as a function of unit characteristics, like the within-KOC models predicting achievement as a function of SES in Massachusetts. Fourth, whatever background characteristics were used in these models would be measured for each of the units in all of the States. Using these data, mean achievement would be predicted for each unit within a State, and aggregated to the State level. Designing a sound model of this kind would require a sophisticated understanding of the demography of the United States. No detailed specification will be attempted here, but a sketch of a possible model can be given.

For units, kinds of communities seem the best choice. School districts are problematical because their average size differs considerably from State to State, because they sometimes overlap, and because background data, e.g. from the decennial United States census, may not be readily aggregated to correspond to school district boundaries. Types of units might be created by crossing the Size and Type of Community (STOC) categories used in the National Assessment of Educational Progress (NAEP) with the four major geographic regions by which NAEP results are reported, and then collapsing across some or all regions within STOC categories where necessary to obtain enough units of a given type to fit a model. Parts of New York, Chicago, and Los Angeles would probably be placed in the same category, for example, even though they are located in different geographic regions, but southern small places might be distinguished from northeastern small places.

Sufficient data should be readily available from the union of NAEP samples in participating States to fit regression models predicting achievement within each type of unit. The predictors used would be limited to data available for all such units in the United States, not just those sampled. In practice, this would probably limit them to data from the United States Census. Priority would be given to educational and economic indicators of socioeconomic level. Each type of unit would have its own regression equation, but for simplicity, it would seem best to use the same set of predictors for all types of units. Nonetheless, there is no technical reason that different sets of predictors could not be used for different types of units. (In the actual construction of the model, the stages of defining unit types and modeling achievement within types would probably be done jointly, not seriatim as presented here. If two provisional community types had similar regression equations, they could be pooled, for example.)

The regression equations for each community type together with background data on all communities would permit the estimation of predicted scores for each unit within each State. In order to aggregate these to the State level, information on the numbers of students in each community would be required. Assuming that schools could be mapped onto communities, this information might come from Quality Education Data (QED) tapes used by Westat to draw NAEP samples, or from the Common Core of Data, or from State education agencies. It would seem simplest if possible to obtain the information from a centralized source.

Given predicted score levels for each community and weights necessary to aggregate those estimates to the State level, predicted score levels for each State would be calculated. Standard formulas from sampling theory would provide acceptable approximations to standard errors at the State level, which could be used to create comparison bands[5]. The primary concern in creating comparison bands would be fairness to each State. Clearly, the sizes of comparison bands would differ from one State to another. In general, the margin of error would be inversely related to the size of the State, for example.

Floating Comparison Groups for States

The method of floating comparison groups used in California might be used for State comparisons with little change. Such a proposal is spelled out in some detail in the March 1988 report of the National Assessment Planning Project conducted by the Council of Chief State School Officers (CCSSO, 1988). Although the CCSSO report recommends further study before the set of variables defining an SES continuum is selected, they suggest provisionally that per capita income, percent of adults having completed 4 years of high school, and percent of school-age children in poverty might be used. States could be ranked according to an equally weighted composite of these three indicators, reversing the direction of percent of children in poverty. Each State could then be compared to a ten-State group, including the five States above and below it as determined by the ranking. The top and bottom five States would each be compared to the same extreme group of ten States. This procedure has the advantage of being simple to understand, and makes good use of the limited number of States available by including each State in five or more different comparison groups. Little is said in the CCSSO report about the actual reporting of the comparison between a given State and its comparison group, but several options might be considered. Most simply, the rank ordering of the State among those in its group could be reported. Still better might be to report only whether each State fell below, within, or above the middle-half of its comparison group. Alternatively, each State's mean could be contrasted with an unweighted average of the achievement means for its comparison group States. (This latter approach would probably yield a series of "expected achievement levels" that were not rank ordered in exactly the same way as the States were. Anomalies might be difficult to explain.)

Summary

There is no wholly satisfactory solution to the problem of presenting State-to-State comparisons fairly, but methods used for interschool and interdistrict comparisons within selected States provide useful starting points. Whatever method is used should be no more than an adjunct to the reporting of unadjusted achievement levels. Models giving predicted levels or providing focused comparison groups are probably best restricted to State-level achievement means, but overall means alone are insufficient for reporting each State's achievement. Additional distributional summaries such as selected quantiles should also be reported, and if feasible, achievement should also be reported for significant student subpopulations.

After reviewing several models used in selected States, three possible approaches were recommended for State-to-State comparisons. First was the reporting of State achievement without any kind of adjustment or clustering, including performance levels for subgroups of students and encouraging comparisons of like students across States rather than gross State-level comparisons. Second was an approach for deriving State achievement comparison bands based on a series of regression models for different types of communities within each major geographic region. Third was the use of floating comparison bands or floating clusters, following the model used in California as proposed in the 1988 CCSSO report.

Footnotes

[1]The term "units" is used throughout to refer to schools, districts, or States--whichever are being compared to one another.

[2]The distribution of school means is not the same as the distribution of individual scores, nor is there in general any simple relationship between the quantiles of the two distributions. Separate norms are required for locating a school in the distribution of schools versus locating a student in the distribution of students. A school's percentile rank cannot be derived from the percentile ranks of its students. This same principle applies to any two or more levels of aggregation, e.g., districts or States. Note that in general, the higher the level of aggregation (i.e., the larger the units compared), the smaller the variance among the means of those units.

[3]Matrix sampling is a procedure under which different students respond to different sets of test items. This permits greatly expanded sampling from the content of the curriculum, because far more items can be used than could practically be given to any one individual. By greatly improving the reliability of school-level means, matrix sampling can dramatically increase the correlations between school achievement scores and predictors like the average parent educational level. The precision of the predicted achievement levels is thereby increased accordingly. With regard to the regression procedure itself, however, it makes no difference whether school

means are derived from an assessment that uses matrix sampling or from a single test given to all students.

[4]Residual scores were obtained by regressing science achievement on the four background factors, then subtracting the predicted science achievement score from the observed score. Science percentiles were obtained within floating comparison groups as described in the text. The high correlation between residuals and floating percentiles indicates that the two procedures are operating in much the same way to control for SES. Scores that appear exceptional under one approach are likely to appear exceptional under the other approach as well.

[5]Note that because of incomplete model specification, errors in regression models would probably be positively correlated across communities of the same kind within a State. These correlations would be difficult to estimate, and would have the effect of increasing the actual standard errors of State-level estimates. In other words, standard errors calculated by ignoring these correlated errors would be too small. This problem would best be minimized by specifying sufficiently homogeneous community types in the first stage of the procedure. Assuming the problem was about as serious in one State as another, it would not seriously compromise the fairness of the procedure to different States.

References

Council of Chief State School Officers. (1988, March). On reporting student achievement at the State level by the National Assessment of Educational Progress (Report of the National Assessment Planning Project, Wilmer S. Cody, director.). Washington, DC: Author.

Fetler, M. (1988, February). Tailored school norms: A method for construction of floating comparison groups. Unpublished manuscript.

Massachusetts Department of Education. (1985). A new classification scheme for communities in Massachusetts. (Publication No.14253-1500-11-85-CR)

Massachusetts Department of Education. (1986). Educational Assessment Report (Sample report for the hypothetical Seaview Middle School, Seaview District, November 1986).

Pennsylvania Department of Education. (1987, May). Educational Quality Assessment: Commentary. Harrisburg, PA: Author.

United States
Department of Education
Washington, DC 20208–5653

Official Business
Penalty for Private Use, $300

POSTAGE AND FEES PAID
U S DEPARTMENT OF EDUCATION
ED 395
FOURTH CLASS BOOK RATE